Human head pose and emotion analysis

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

Aryaman Gupta
201302087
aryaman.gupta@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
March 2021

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Human head pose and emotion analysis" by Aryaman Gupta, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Adviser: Prof. Vineet Gandhi

To Patience

Acknowledgments

As I submit my thesis, I extend my gratitude to all those people who helped me in successfully completing this journey. Foremost my deepest gratitude to Prof. Vineet Gandhi for his constant support and encouragement to explore and chart my own research path. I would also show my gratitude to the CVIT ecosystems, through which I came to have multiple technical/non-technical discussions with wide background of students in PhD/MS/Honors and also developed an understanding of the vast variety of problems begin attempted at the centre.

I will also like to thank my parents whose support has been always immense in all my adventures in life. I would like to thank Kalpit Thakkar and KL Bhanu for their advice and support. I would also like to mention Abhijeet Kumar as a lab partner and a friend for the important discussions. I would also like to thank Ashutosh Sharma for his constant motivation.

Abstract

Scene analysis has been a topic of great interest in computer vision. Humans are the most important and most complex subject involved in scene analysis. Humans exhibit different forms of expressions and behaviour with its environment. These interactions with its environment have been in study for a long time and to interpret these interactions various challenges and tasks have been identified.

We focus on two tasks in particular: Head Pose estimation and Emotion recognition. Head poses are an important mean of non-verbal human communication and thus a crucial element in understanding human interaction with its environment. Head pose estimation allows a robot to estimate the region of focus of attention for an individual. Head pose estimation requires learning a model that computes the intrinsic Euler angles for pose (yaw, pitch, roll) from an input image of the human face. Annotating ground truth head pose angles for images in the wild is difficult and requires ad-hoc fitting procedures (which provides only coarse and approximate annotations). This highlights the need for approaches which can train on data captured in a controlled environment and generalize on the images in the wild (with varying appearance and illumination of the face). Most present day deep learning approaches which learn a regression function directly on the input images fail to do so. To this end, we propose to use a higher level representation to regress the head pose while using deep learning architectures. More specifically, we use the uncertainty maps in the form of 2D soft localization heatmap images over five facial keypoints, namely left ear, right ear, left eye, right eye and nose, and pass them through a convolutional neural network to regress the head-pose. We show head pose estimation results on two challenging benchmarks BIWI and AFLW and our approach surpasses the state of the art on both the datasets. We also propose a synthetically generated dataset for head pose estimation.

Emotions are fundamental to human lives and decision-making. Human emotion detection can be helpful in understanding human mood, intent or choice of action. Recognizing emotions from images or video accurately is not easy for humans themselves and for machines it is even more challenging as humans express their emotions in different forms and there is a lack of temporal boundaries among emotions. Facial Expression Recognition has remained a challenging and interesting problem in computer vision. Despite efforts made in developing various methods for facial expression recognition, existing approaches lack generalizability when applied to unseen images or those that are captured in wild setting (i.e. the results are not significant). We propose use of facial action unit's soft localization heatmap images for facial expression recognition. To account for lack of large well labelled dataset we propose

a method for automated spectrogram annotation where we use two modalities(visual and textual) used in expression of emotion by humans to label one other modality(speech) for emotion recognition.

Contents

Chapter Page

1 Introduction . . . 1
1.1 Head pose estimation . . . 1
1.2 Human Emotion recognition in scenes . . . 2
1.3 Associated Challenges: . . . 3
1.3.1 Head pose estimation challenges . . . 3
1.3.2 Human emotion recognition challenges . . . 4
1.4 Our contributions . . . 4
1.4.1 Head Pose Estimation contributions . . . 4
1.4.2 Human emotion recognition contributions . . . 4
1.5 Thesis Overview . . . 5

2 Related Work . . . 6
2.1 Head pose estimation . . . 6
2.2 Human emotion recognition . . . 8

3 Head Pose Estimation . . . 10
3.1 Method for keypoints uncertainty maps . . . 10
3.2 Proposed approach . . . 12
3.3 Experiments and results . . . 14
3.3.1 Experimental Setup and Datasets . . . 14
3.3.2 Results . . . 14
3.4 Synthetically generated dataset . . . 16
3.4.1 Generation of synthetic samples . . . 17
3.4.2 Experiments and results . . . 18

4 Human emotion recognition . . . 21
4.1 Facial expression recognition using action units heatmap . . . 21
4.1.1 Facial action units heatmaps generation . . . 22
4.1.2 Method . . . 23
4.1.3 Experiments and results . . . 24
4.2 Automated spectrogram annotation for emotion recognition . . . 25
4.2.1 Proposed approach . . . 26
4.2.1.1 Automated spectrogram annotation . . . 26
4.2.1.2 Training using spectrogram . . . 27
4.2.2 Experiments and results . . . 27

5 Conclusions 29

Bibliography 31

List of Figures

Figure Page

1.1 Axis of rotation for head: yaw, pitch and roll 2
1.2 Different facial expressions 3

2.1 CNN architecture used by Ruiz et. al[56] for head pose estimation 7
2.2 Overview of recent approaches for facial expression recognition using deep neural networks 9

3.1 Convolutional Pose Machine [74] comprises of several stages in consecutive manner sequence in order to make dense prediction of each image location. In the figure we show the prediction improving at each stage for the location of the right elbow (a) Prediction based on local context is less accurate. (b) Context from different parts helps to resolve ambiguity (c) More iterations help converge to a certain solution 10
3.2 Framework and receptive fields of CPMs. The figure demonstrates the CNN architecture and receptive field across layes for a CPM with T stages At each subsequent stage the receptive field size increases helping to get cues from more parts and thus helping the prediction getting more accurate. In the first stage(a) the the architecture works just on the RGB image. Insets (b) and (d) illustrates that in the subsequent stages, the network works on both image as well as belief maps generated from preceding stage. Inset (e) we show the effective receptive field on an image (centered at left knee) of the architecture, where the large receptive field enables the model to capture long-range spatial dependencies such as those between head and knees 11
3.3 Spatial context from easier to locate part's belief map can help in giving strong cues for localizing difficult to locate part's belief map. In this example, spatial contexts from shoulder, neck and head can help eliminate wrong (red) and strengthen correct (green) estimations on the belief map of right elbow in the subsequent stages. 12
3.4 Example of a face image, detected keypoints and respective heatmaps of each keypoint computed using [8]. 13
3.5 The architecture consists of 3 convolutional layers (conv1, conv2, conv3) followed by two fully connected layers (fc1, fc2). The input has 5 channels: one each for the nose, left eye, right eye, left ear and right ear (heatmap images for these keypoints). The network outputs the estimated values of the three intrinsic Euler angles (yaw, pitch, roll). 13
3.6 Samples from biwi head pose dataset 15
3.7 Samples from Aflw dataset 17

3.8 Estimation of head pose using three different models (all trained on BIWI), on unseen images taken from the web. **Top row**: Results for CNN-based model [82] which takes RGB images as input, **Bottom row**: Results for our CNN-based framework which takes heatmaps of five facial keypoints locations as input. 18
3.9 Uneven distribution in AFLW dataset(consists of real-world head images) across each angle value, respectively, i.e., pitch, yaw, and roll 19
3.10 One of the 3d face model and some of the samples generated from it by rotation 19
3.11 Samples from Nvidia Synthetic Head Pose dataset[19].**Top row**: Different face models, **Bottom row**: Head poses for top right face model 20

4.1 Target heatmaps for a particular representative image. The peak and size of the heatmaps are specified by the corresponding labels, and are placed based on landmarks which help in locating Action Unit. These heatmaps are concatenated to construct the heatmap regression. 22
4.2 The image is processed through the Hourglass network, generating the heatmaps corresponding to the Action Units, at the location they occupy. The heatmaps are activated based on the predicted intensity. 22
4.3 CNN architecture used for the facial emotion recognition. The 8 channel input conists of 3 RGB channels and 5 AU heatmap channels 23
4.4 Six sample images from CK+ database 24
4.5 Four sample images from JAFFE database with different emotion on the same subject . 24
4.6 Steps in our pipeline for the process of automated annotation of video clips based on emotion 26
4.7 Example spectrogram 27

List of Tables

Table Page

3.1 Results on BIWI with 8-fold cross-validation (21 randomly selected videos for training and the remaining 3 videos for test such that no person appears both in training and test sets) . . . 15
3.2 Results on AFLW dataset with 5-fold cross validation. *: Constrains the angles to a certain range. . . . 16
3.3 Results on AFLW using testing protocol in [12]. . . . 16
3.4 Results with our method using model trained exclusively with our generated synthetic head pose dataset . . . 18
3.5 Results on Nvidia Synthetic Head Pose dataset[19] . . . 18

4.1 Average accuracy on the CK+ database for seven expressions classification with 6-fold cross validation . . . 25
4.2 Average accuracy on JAFFE with 6-fold cross validation . . . 25
4.3 Distribution of samples in our database . . . 28

Chapter 1

Introduction

Human faces are very essential to understand human behaviour. Human face perception[73] helps in understanding and interpretation associated information processing in the human mind. Facial features carry a wealth of social information[14]. For a machine to understand human behaviour, it has to inference the human face visual signal very accurately, quickly and robustly irrespective of the occlusion, awkward angles, lighting, etc. The two very important aspects to human faces are the expression and its head orientation. Head pose orientation can be useful for determining human's visual focus region. Human emotions facilitate adaptive responses to environmental challenges.

Face images have been used to train models for these tasks in most of the methods directly. With recent developments in deep learning these models have performed well over datasets which are captured in controlled environments. But the models learned are not very robust for the tasks, they do not predict quite accurately for samples from different environments as the model learns facial feature components which are not entirely relevant to the tasks. Moreover people with different facial structure and expressive styles create unique challenge in separating the relevant features. We approach it in a way where we first predict task specific uncertainty maps and use those uncertainty maps to train our model. We identify there is inherent uncertainty in predicting these uncertainty maps. To account for this uncertainty in localization we represent them using uncertainty maps (heatmap images) which capture the soft localization of these features.

In deep learning research, a large well labelled dataset is very crucial for training the model. We propose method to obtain well labelled dataset using an automated process to help in overcoming manual annotation cost.

1.1 Head pose estimation

Head pose estimation requires learning a model that computes the intrinsic Euler angles for pose (yaw, pitch, roll) from an input image of the human face. In order to humanize machines by bringing them closer to human-like perception and under- standing, accurately estimating the human head orientation using visual imagery presents an important challenge. Head pose inferences visual attention

and interest of a person, which is crucial for many applications in computer vision. Estimating head pose has been actively pursued in problems like social event analysis [1], Human Computer Interaction (HCI) [2], driver assistance systems [3] etc., which are an important part of present day technologies. Formally, head pose estimation entails computing the 3D orientation of head with respect to the camera pose using digital images. Initial approaches estimated only one or two angles for head pose while assuming other angles are fixed or fixed discrete values for head pose angles to be estimated.

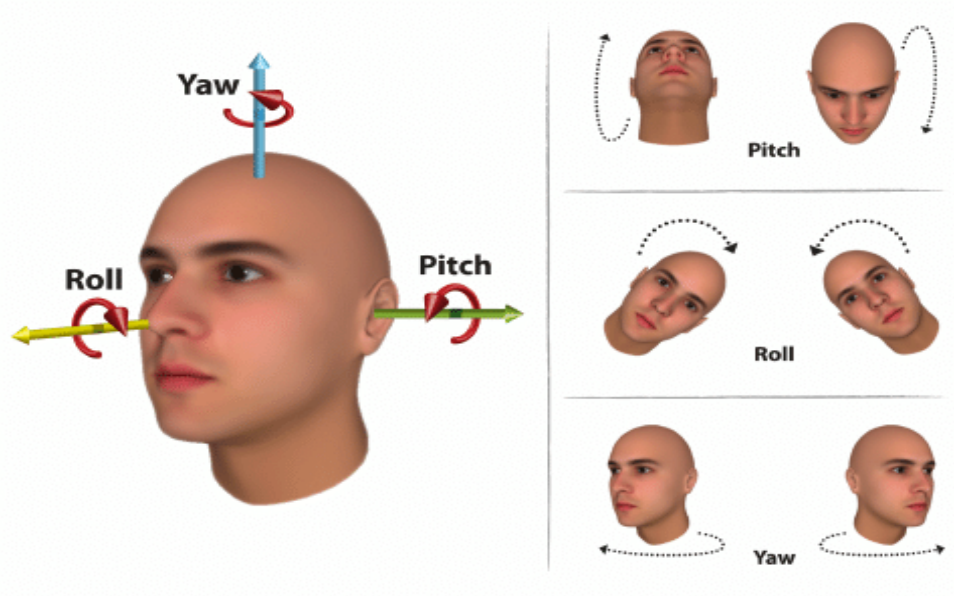

Figure 1.1 Axis of rotation for head: yaw, pitch and roll

While computer vision based pose estimation approaches have focused predominantly on appearance-based solutions that compute human pose directly from digital images, there have been methods based on psycho-physical experiments. These consider the human perception of head pose to rely on cues such as deviation of nose angle and the deviation of the head from bilateral symmetry [78]. Since it is easier to annotate 2D keypoints directly on images, huge labelled datasets for keypoints are now available [1] and have lead to development of powerful methods [8] for localizing keypoints like nose, eyes and ears. We hypothesize that we can learn a head pose estimation model using only five facial keypoint locations. Such a model implicates an abstraction over the appearance and illumination dependent image data which is a hindrance for generalization capability of head pose estimation methods. The abstraction limits the dependencies of the model to scale and configuration of a few keypoint locations.

1.2 Human Emotion recognition in scenes

Human emotion plays a significant role [84] in social interaction, human intelligence, environment perception, human intent etc. Automatically human emotion recognition helps improve human-

computer interactions. The task of recognising emotions is challenging for humans themselves because human emotions lack distinguishing boundaries and different individuals express emotions in different ways. Humans use verbal and of non-verbal cues, such as facial expressions, gestures, body language and tone of voice, to express their emotions. For computers, emotion recognition is an even more challenging task.

Figure 1.2 Different facial expressions

For facial emotion recognition, CNN that use faces images to train the model [29] achieve a high accuracy in emotion recognition. We from our experience with head pose estimation use action units uncertainty maps for training our model.

Obtaining large, human labelled datasets to train models for emotion recognition is a notoriously challenging task, hindered by annotation cost and label ambiguity. We try to solve this problem by generating a labelled dataset for emotion recognition in automated way.

1.3 Associated Challenges:

1.3.1 Head pose estimation challenges

- Occlusion: There are generally multiple objects in the image, some of them occlude our focus area of faces. For ex. shades, cap, jewellery, mask etc. Occlusion creates uncertainty in determining the location of the keypoints which are not visible.

- Extreme head orientations: Some of the face images have little or no visibility of some major keypoints in extreme head orientation

- Dataset and annotation: BIWI captured in a constrained environment while AFLW annotation are less precise than BIWI
- Facial structure variations: Different people have different face structure thus the relative spatial arrangement for the keypoints is not same for different individuals, this poses a challenge to train a model based on spatial arrangement of keypoints

1.3.2 Human emotion recognition challenges

- Annotation: Emotions lack distinguishable boundaries and different individuals express emotions in different ways making labelling very subjective
- Occlusion: There are generally multiple objects in the image, some of them occlude our focus area of faces. For ex. shades, cap, jewellery, mask etc. Occlusion creates uncertainty in determining features which are not visible.

1.4 Our contributions

1.4.1 Head Pose Estimation contributions

- A hypothesis on learning a model for head pose estimation which relies only on five facial keypoint locations and abstracts out the dependency on appearance of the subject.
- A baseline approach that uses the exact keypoint locations (sampled from their distribution) and employs a MLP for regression of pose angles.
- A CNN-based framework which uses the probability distribution of keypoint locations in the form of heatmap images, as input to regress the head pose.
- State-of-the-art performance for head pose estimation using the CNN-based framework on the BIWI [13] and AFLW [40] datasets.
- A more robust approach to train model using synthetically generated dataset

1.4.2 Human emotion recognition contributions

- A CNN-based framework which uses the probability distribution of facial action units as input to learn model for facial emotion recognition, which achieves close to state of the art performance on CK+ [39] and Jaffe [28] datasets
- An automated process to generate annotated spectrogram for emotion labels from feature length movie videos and collection of a corresponding dataset for speech emotion recognition

1.5 Thesis Overview

Now we provide the organizational scheme of the upcoming chapters in this thesis. Chapter 2 provides the literature survey for head pose estimation and emotion recognition. In first section of Chapter 3, we detail out our approach for head pose estimation using which we achieve state of the art results on head pose estimation. In second section of chapter 3 we describe how synthetic dataset for head pose estimation can be effective for training using our method for head pose estimation and also our method for creating synthetic dataset using 3D head models. In first section of Chapter 4 describes our approach, experiments and results for facial emotion recognition. In second section of chapter 4 we dive into our method for generation of annotated spectrogram from feature length movie videos. Finally chapter 5 summarizes the main contributions of the thesis.

Chapter 2

Related Work

2.1 Head pose estimation

Previous approaches to head pose estimation can be classified into two categories: RGB and RGBD based (2D vs 3D input). We limit our discussion to RGB input only. Earlier methods for head pose estimation used appearance templates that use a set of exemplars to find the pose of an input image, by finding the closest exemplar [26]. The assumption that similarity in image space equates similarity in pose is the major drawback of such methods. Extending appearance templates, several methods using multiple pose detectors (each corresponding to one discrete pose) have been proposed [31]. However, detector-based methods require several detectors and non-face samples (negative samples) for successful training, which is burdensome. Manifold embedding methods were later introduced, which project an input sample to a lower dimension using an embedding function and regress pose in the embedding space. Techniques like PCA [79], Isomap [25] and several combinations [81] of dimensionality reduction approaches are used for head pose estimation. Learning useful low-dimensional representations needs proper training data having balanced samples.

With the transition to deep learning based methods, several former drawbacks have been mitigated. One of the earliest efforts in this area was by Osadchy et. al [51]. They extract CNN features from images and regress pose using them. Patacchiola and Cangelosi [52] test the effect of dropout and adaptive gradient-based methods combined with CNNs for head pose estimation, where they propose to use adaptive gradients in conjunction with a CNN. On the other hand, Ruiz et. al [56] propose a CNN with 3 separate branches, each with combined classification and regression for the respective head pose angle. Both these methods aim to improve performance of head pose estimation in the wild. Lathuiliére et. al [32] proposed a CNN-based model with a Gaussian mixture of linear inverse regressions. They use an Imagenet-pretrained CNN to learn face features and train a pose regressor on them. An extension of this approach by Drouard et. al [11] proposes to cope with changing illumination conditions, variability in face orientation and in appearance, etc. by combining the qualities of unsupervised manifold learning and inverse regressions. However, as the CNN-based methods estimate the pose angles directly from RGB images, it makes them prone to poor generalization on account of illumination as well as appear-

ance changes. The architecture used by Ruiz et. al[56] for training model directly from face images has been depicted in figure 2.1

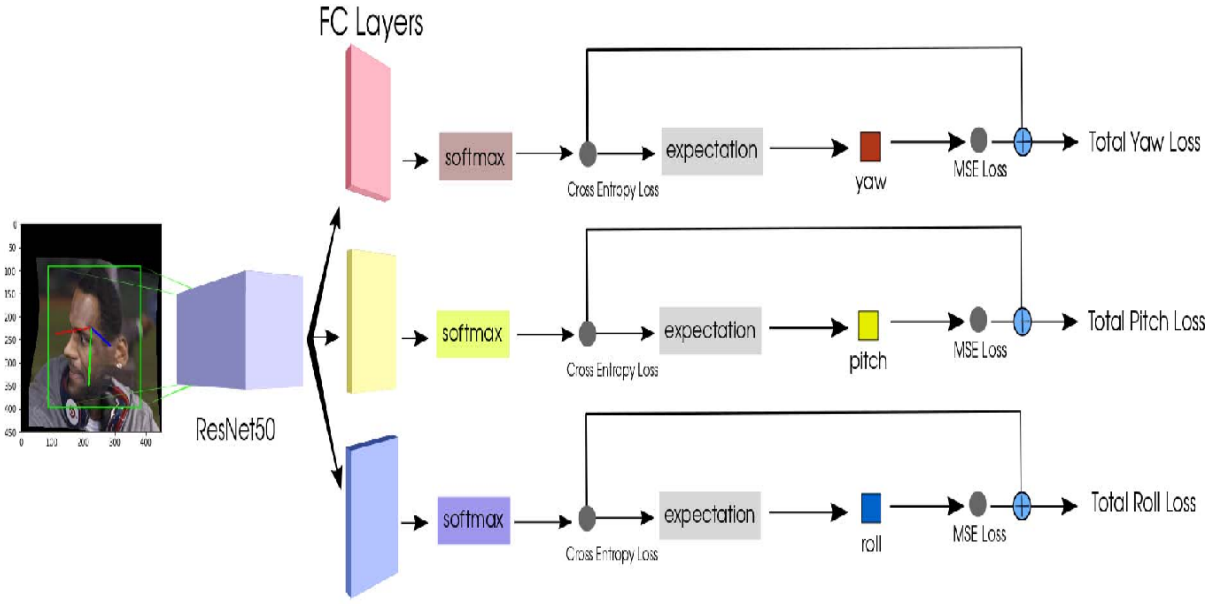

Figure 2.1 CNN architecture used by Ruiz et. al[56] for head pose estimation

Geometric models regress the pose using facial features such as keypoints, nose angle, etc. and have been proposed in previous literature [71]. Similar in spirit, we propose to use a higher-level feature to drive the pose regression, viz. the heatmaps of five facial keypoints extracted from face images (or exact 2D locations) using a keypoint localization routine [8]. The performance of our models prove our hypothesis of facilitating abstraction over illumination and appearance dependent image data by achieving state-of-the-art results for head pose estimation and demonstrating good generalization capability.

In the past decades, researchers have made impressive progress on 3D object modeling and synthesis. Synthesized data has been applied for deep network training in many computer graphics and vision tasks, e.g., autonomous driving, license plate recognition 3D reconstruction, scene understanding, human detection and pose estimation , and medical image segmentation and detection, which have proved that synthetic data can help to achieve good performance. We find that the head pose datasets (e.g,. Biwi Kinect Head Pose Dataset [13] and Pointing'04 Dataset [18]) are not sufficient for training a deep network due to the limited variation, the number of available samples, and partially incomplete annotation.

Deep learning methods for the task of head pose estimation rely heavily on large well annotated data. Creating a large and precisely annotated dataset for the task of head pose estimation with the help of manual annotation is quite challenging. Synthetic dataset created using images of rendered 3D face can thus get rid of this problem. Gu et al. [19] propose a synthetic head pose dataset, SynHead. Their method is based on sequential prediction with the use of recurrent neural networks. Their model is trained using synthetic dataset and fine tuned using real world dataset. It is hard to use SynHead dataset

for monocular head pose estimation as it has been created for sequential head pose estimation. Liu et al. [36] also used their own synthetic head pose dataset for training model. They use cross-dataset testing, by training their model on synthetic dataset and testing on real-world dataset. They haven't made their dataset publicly available.

2.2 Human emotion recognition

In the earliest approaches on emotion recognition, most common method was a two step machine learning approach, the first involves obtaining features from the face images and in the second step, machine learning models are used to classify images based on the extracted features. The feature descriptors used by conventional methods include Haar features[76], histogram of oriented gradients (HOG) [9], Local Gabor features[3] and local binary patterns (LBP) [57]. Based on these features, classifier associates emotion to the face image. The approaches performed fairly well on earlier datasets but with rise of more complex datasets (having increased intra class distinction) lack of robustness of these methods implied poor performance on these new datasets.

Deep learning has been widely successful in image classification and other vision problems [30], [23], [38], [45], [46], [17]. For the task of facial expression recognition(FER) many groups have successfully deployed deep learning. The work of Khorrami et al [29] demonstrated high accuracy in facial emotion recognition can be attained with the use of CNN on the extended Cohn-Kanade dataset (CK+)[39]. The method proposed by Mollahosseini [48] used neural network for facial expression recognition using 2 convolution layers, 1 max pooling layer, and 4 "inception" layers, i.e. sub-networks. Identity-aware CNN (IA-CNN) introduced by Meng et al[43] uses expression sensitive and identity contrastive losses to diminish the contrast in expression-related information and learning identity respectively. A review of deep learning approaches in recent years for facial expression recognition is showcased in figure 2.2

Facial action unit[83] are used to identify human facial movements by their appearance on the face. Using combinations of action unit(AU), facial expression can be inferred. As AUs are independent of any interpretation, they can be used for any higher order decision making process including recognition of basic emotions, or pre-programmed commands for an ambient intelligent environment. AUs are a contraction or relaxation of one or more muscles. AU detection has been studied for decades and various methods have been proposed [41] for emotion recognition using it. To achieve good performance, researchers have designed different features to represent AU. The features include the appearance texture of the whole face [88] or near the facial landmarks [70, 72], or the combination of geometry shape with texture [5]. Most of these features are based on general features in computer vision task, such as SIFT, HOG, LBP, etc. To make the features discriminative for AU, some works considered that AU is tightly correlated to the motions within local regions of the face [15] and thus introduced sparsity-induced algorithms [89, 63] to reduce the influence of uncorrelated facial regions. Over the last few years, deep learning has become a dominating approach due to their capability and capacity of representation

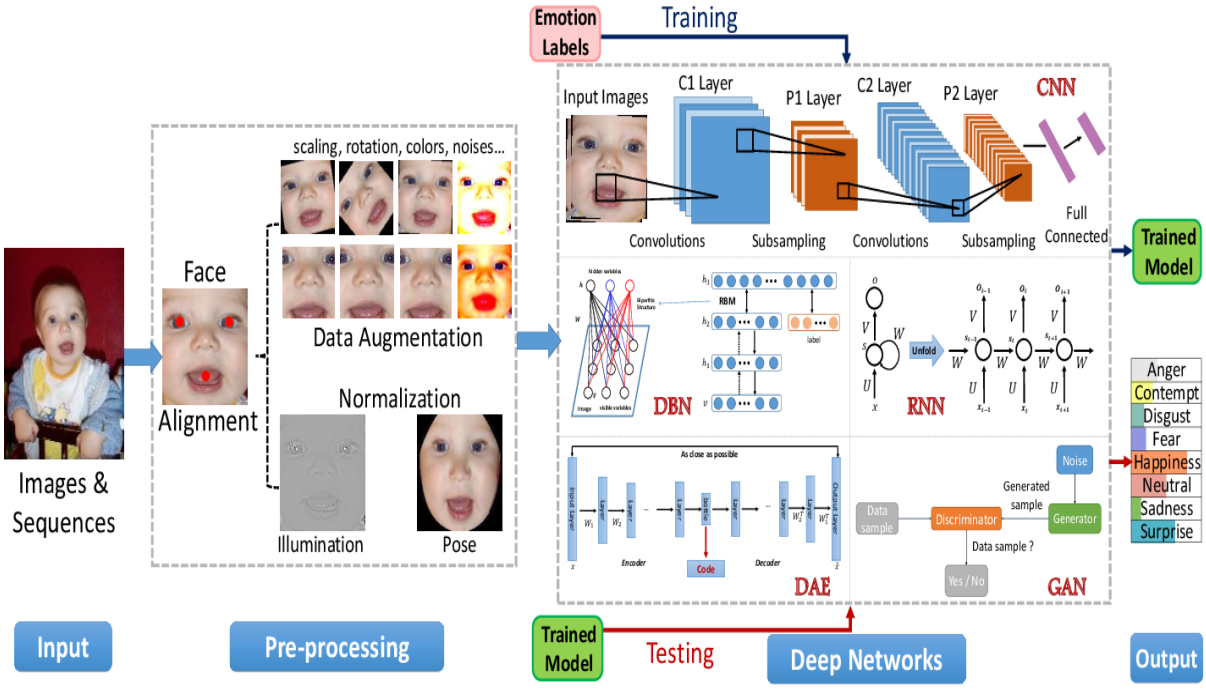

Figure 2.2 Overview of recent approaches for facial expression recognition using deep neural networks

learning. These methods [91, 34] learn rich local features to capture facial deformation. To mention some, Zhao et al. [91] introduced a locally connected CNN layer that learns region specific CNN filters from parts of the face. JPML [89], EAC-Net [34] and ROI [33] obtained features for facial landmarks that are robust with respect to non-rigid shape changes. JAA-Net[58] proposed to jointly learn AU detection and face alignment in a unified framework. These methods have achieved promising performance on annotated datasets, e.g., CK+ [39], DISFA [42], BP4D [86]. However, these approaches rely on precisely labelled face images and often overfit on a particular dataset because of inadequate amount of training data. To reduce the reliance of AUs annotations, few approaches move to learning model in a semi-supervised [85, 10], weakly-supervised [90, 53, 87] or self-supervised manner [77]. The semi-supervised learning methods usually comprise both unlabelled and labelled data by assuming the faces to be clustered by AUs, or to have a smooth label space. The weakly supervised approaches exploit noisy, incomplete AU annotations [90]. They usually learn AU classifiers from domain knowledge [87], or naturally existing constraints on AUs [53]. We adopt the self-supervised learning paradigm because it can learn AU discriminative features without AU labels and it is regardless of the assumptions on label distribution.

Humans express their emotion using different forms. Apart from the work on facial expression recognition as mentioned above, there are several works[47, 35] based on speech processing. More recently, other modalities such as body gestures, text have been started to be used. Approaches for sentiment analysis and emotion recognition using multi modality[55, 67] has also come into prominence.

Chapter 3

Head Pose Estimation

Recently, work has developed on estimating head pose using neural networks. We propose a method where we compute uncertainty maps for our five face keypoints. We use those uncertainty maps as input for the training our model for head pose estimation. While there are various methods for computing keypoints, we prefer to obtain uncertainty maps from the method described in [75] and [8]. The method is able to predict keypoints even in case of complete occlusion of keypoints as well as in case of extreme head orientation rotation angles. This method's ability to overcome these challenges for keypoints prediction is quite crucial to our method's robust performance in challenging samples for head pose estimation. We describe this method for keypoints prediction in the next section.

3.1 Method for keypoints uncertainty maps

Figure 3.1 Convolutional Pose Machine [74] comprises of several stages in consecutive manner sequence in order to make dense prediction of each image location. In the figure we show the prediction improving at each stage for the location of the right elbow (a) Prediction based on local context is less accurate. (b) Context from different parts helps to resolve ambiguity (c) More iterations help converge to a certain solution

For the task of keypoints prediction Convolutional Pose Machines (CPMs) are used. CPMs inherit the advantages of the pose machine [68] architecture—the implicit learning of long-range dependencies between image and multi-part cues, tight integration between learning and inference, a modular sequential design—and combine them with the benefits afforded by convolutional architectures: the power to find out feature representations for both image and spatial context directly from data; a differentiable architecture that permits for globally joint training with backpropagation; and therefore the ability to efficiently handle large training datasets.

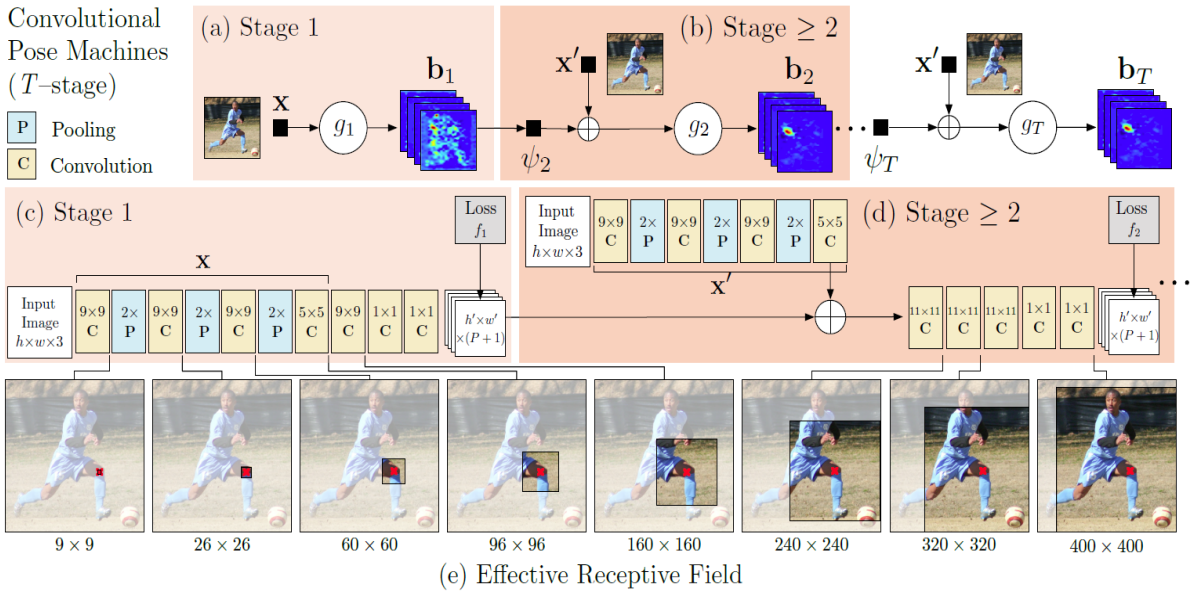

Figure 3.2 Framework and receptive fields of CPMs. The figure demonstrates the CNN architecture and receptive field across layes for a CPM with T stages At each subsequent stage the receptive field size increases helping to get cues from more parts and thus helping the prediction getting more accurate. In the first stage(a) the the architecture works just on the RGB image. Insets (b) and (d) illustrates that in the subsequent stages, the network works on both image as well as belief maps generated from preceding stage. Inset (e) we show the effective receptive field on an image (centered at left knee) of the architecture, where the large receptive field enables the model to capture long-range spatial dependencies such as those between head and knees

There are multiple stages in CPMs [74] comprised of 2D belief maps for the location of each part. In each of CPM stage, image features and the belief maps produced by the previous stage are used as input. The belief maps provide the subsequent stage an expressive non-parametric encoding of the spatial uncertainty of location for each part, allowing the CPM to learn rich image-dependent spatial models of the relationships between parts. Instead of explicitly parsing such belief maps either using graphical models [54, 64, 65] or specialized post-processing steps [64, 66],it trains convolutional networks that legitimately work on moderate conviction maps and learn verifiable picture subordinate spatial models of the connections between parts. The general proposed multistage engineering is completely differentiable and in this way can be prepared in a start to finish design utilizing backpropagation. At a specific

stage in the CPM, the spatial setting of part convictions give solid disambiguating signals to an ensuing stage. Thus, at every subsequent stage of CPM, it produces belief maps with increasingly refined estimates for the locations of each part (see Figure 3.1). To catch long reach spatial relations between parts, the plan of the organization in each phase of our successive forecast system is propelled by the objective of accomplishing an enormous open field on both the image and the belief maps. It is discovered through investigations, that enormous open fields on the belief maps are critical for learning long reach spatial connections and result in improved accuracy.

CPM is comprised of several convolutional networks, thus the overall network consists of many layers which can lead to issue of vanishing gradients [4, 2, 16, 24] during learning. This issue can happen in light of the fact that back-propagated gradients diminish in strength as they are propagated through the many layers of the network. While there exists recent work which shows that supervising very deep networks at intermediate layers aids in learning [7, 61], they have mostly been restricted to classification problems. Pose estimation is a structured prediction problem, CPMs naturally suggest a systematic framework that replenishes gradients and guides the network to produce increasingly accurate belief maps by enforcing intermediate supervision periodically through the network.

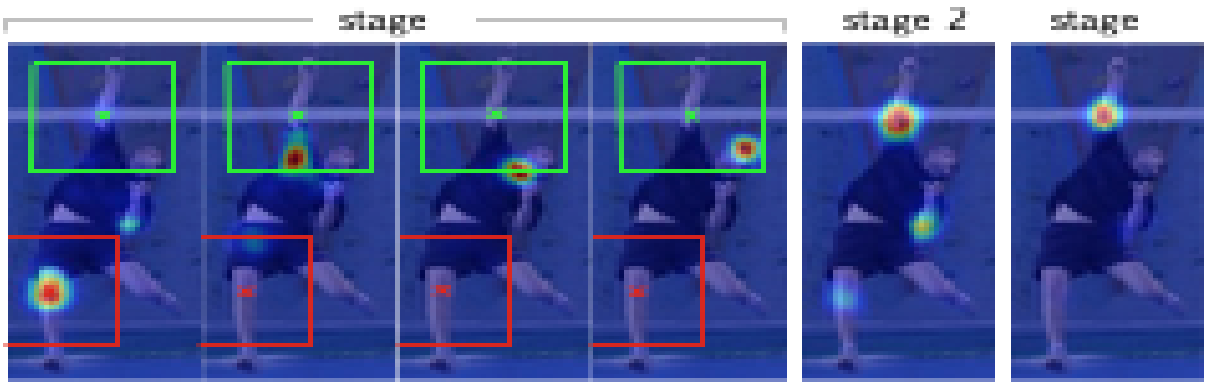

Figure 3.3 Spatial context from easier to locate part's belief map can help in giving strong cues for localizing difficult to locate part's belief map. In this example, spatial contexts from shoulder, neck and head can help eliminate wrong (red) and strengthen correct (green) estimations on the belief map of right elbow in the subsequent stages.

3.2 Proposed approach

Our baseline approach is to employ a Multi Layer Perceptron (MLP) which regresses the 3D head-pose directly using the predicted locations of the five keypoints (detected using [8]). Each of the keypoint is parameterized by its 2D location and prediction likelihood, resulting in an input vector of 15 dimensions, which is used to regress a 3D vector representing the yaw, pitch and roll. Undetected keypoints are represented by a vector of zeroes.

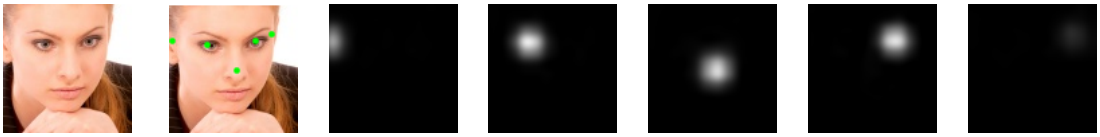

Figure 3.4 Example of a face image, detected keypoints and respective heatmaps of each keypoint computed using [8].

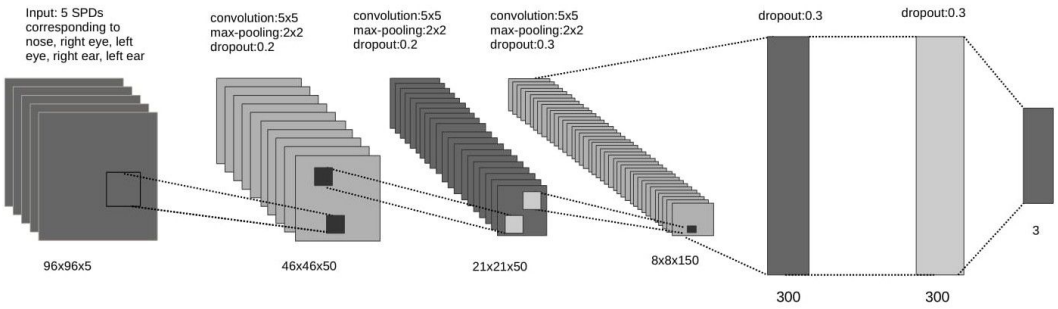

Figure 3.5 The architecture consists of 3 convolutional layers (conv1, conv2, conv3) followed by two fully connected layers (fc1, fc2). The input has 5 channels: one each for the nose, left eye, right eye, left ear and right ear (heatmap images for these keypoints). The network outputs the estimated values of the three intrinsic Euler angles (yaw, pitch, roll).

MLP-based method is based on the assumption that the locations of five facial keypoints estimated from the face image are accurate. However, in practice there is inherent uncertainty in predicting the locations of keypoints such as eyes, ear and nose. One possible way to account for this uncertainty in localization is to treat the image locations of the facial keypoints as latent variables. From a representation perspective, uncertainty maps (heatmap images) can be used to depict latent variables, which capture the soft localization of 2D keypoint locations (Figure 3.4 illustrates an image and corresponding uncertainty maps for the five different facial keypoints used in our work). An image-based representation of the facial keypoint locations facilitates the use of CNN-based approaches for learning the head pose.

Uncertainty maps over locations of keypoints (or joints) in human body or an object skeleton, present in an image, have been successfully used in previous literature where the exact locations of the keypoints were noisy or unknown. Zhou [93] use heatmap images of 2D joint locations to infer 3D human pose using an Expectation Maximization framework. Wu [80] use heatmaps of 2D skeleton keypoints of an object as an intermediate representation to recover 3D structure of an object and bridge the gap between synthetic and real data. Interestingly, both these works [93, 80] use heatmaps over 2D spatial locations to infer 3D structure/pose. Deriving motivation from these efforts, we propose an algorithm which takes 2D uncertainty maps over the facial keypoints as input and regresses the 3D head pose.

Unlike previous efforts [93, 80] that use heatmaps as an intermediate representation and do not have ground truth data, we have ground truth pose angles available. This allows us to directly train a

convolutional regression network using ground truth supervision for head pose estimation. We use five facial keypoint locations as illustrated in Figure 3.4. Each heatmap image is considered as a separate channel and the channels are stacked together, which generates a 5-channel feature map. This feature map is used as an input to the CNN, the architecture of which is shown in Figure 3.5, to learn a head pose estimation model. The final layer gives the values of three pose angles obtained as a result of the convolutional regression. We use a MSE loss to train the convolutional regression network, which can be written as follows:

$$\mathbf{L}_{\text{mse}} = \frac{1}{3}\sum_{i=1}^{3}\left(\mathbf{\Theta}_i - \hat{\mathbf{\Theta}}_i\right)^2 \tag{3.1}$$

where, $\mathbf{\Theta}_i$ is the vector consisting of the predicted values for intrinsic Euler angles and $\hat{\mathbf{\Theta}}_i$ is the vector consisting of the values of ground truth angles.

3.3 Experiments and results

3.3.1 Experimental Setup and Datasets

MLP-based Model Our network consists two hidden layers of size 30 neurons each. We set learning rate of 0.00001 and train for 500 epochs using Adam optimizer with a weight decay of 0.0001 and batch size 64.

CNN-based Model We use a convolutional neural network architecture with three convolution layers and two fully connected layers (we have used same architecture used in Liu[37] but with five input channels). Training is run for 1200 epochs with Adam optimizer and set learning rate of 0.00001. We set the batch size to 32. All the experiments are run on a single Nvidia GTX 1080Ti GPU.

We use two benchmark datasets to measure the performance of our models and test them. **BIWI** Kinect Headpose Dataset [13] contains over 15K samples spread over 24 sequences, captured in a controlled environment. The range of head pose angles in the dataset vary from $\pm 75^\circ$ for yaw, $\pm 60^\circ$ for pitch and $\pm 50^\circ$ for roll. **AFLW** [40] Annotated Facial Landmarks in the Wild (AFLW) provides a large collection of annotated face images gathered from the web, exhibiting a large variety in appearance (e.g., pose, expression, ethnicity, age, gender) as well as general imaging and environmental conditions. In total about 25K faces are annotated with up to 21 landmarks per image.

3.3.2 Results

Results on BIWI dataset: As BIWI is captured in controlled conditions and has better ground truth annotations, better performance is achieved on this dataset. The motivation for designing our frameworks is to train a model on a dataset like BIWI and use it to generalize to face images in the wild. In order to demonstrate the ability of our frameworks, we predict the head pose on unseen images taken from the

Method	Yaw	Pitch	Roll	MAE
Liu [37]	6.0	6.1	5.7	5.94
Ruiz et al. [56]	4.810	6.606	3.269	4.895
Drouard [11]	4.24	5.43	4.13	4.6
DMLIR [32]	**3.12**	4.68	3.07	3.62
MLP with location (Ours)	3.64	4.42	3.19	3.75
CNN + Heatmaps (Ours)	3.46	**3.49**	**2.74**	**3.23**

Table 3.1 Results on BIWI with 8-fold cross-validation (21 randomly selected videos for training and the remaining 3 videos for test such that no person appears both in training and test sets)

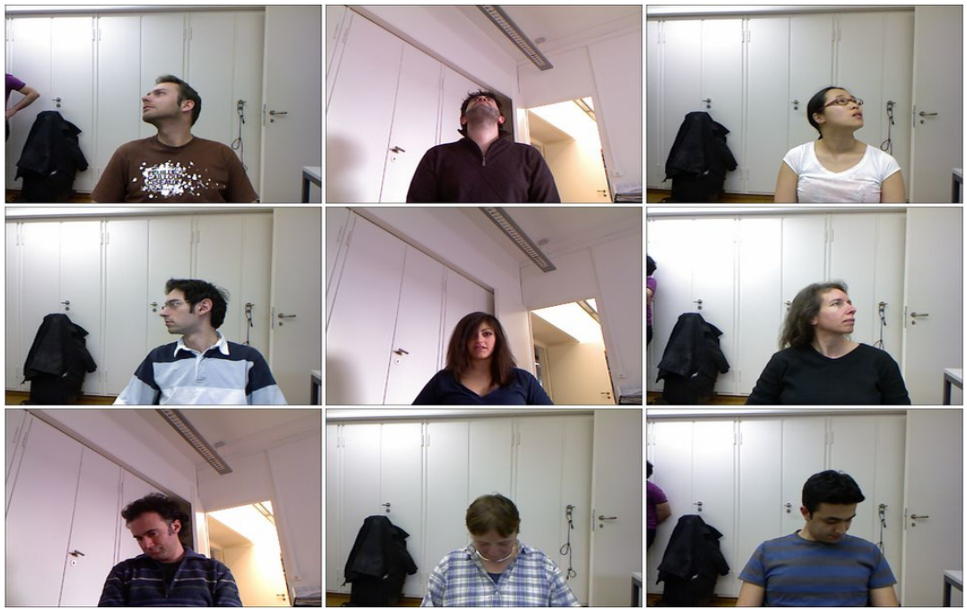

Figure 3.6 Samples from biwi head pose dataset

web (as illustrated in Figure 3.8). Our results show the presence of a perceptually better sense of pose than a model learned directly on the RGB images. Quantitative results for the dataset in terms of Mean Absolute Error (MAE) from ground truth annotations are shown in Table 3.1 which shows that the MLP model achieves competitive performance, while the CNN based approach surpasses the state of the art.

Results on AFLW dataset Given the large variations in AFLW dataset, most of the previous methods compute results for head pose estimation on this dataset by constraining the range of angles, using a subsampled set of images or creating a very small test set [56, 52]. We do not assume any such constraints and show the results using a standard five-fold validation process on the entire dataset, where the samples are randomly divided into train and test sets with 80% samples ending up in training set. We also perform experiment following testing protocol in [12] (i.e. selecting 1000 images from testing and remaining for training) and present the results in Table 3.3. The numbers of other methods in both tables are reported directly from the associated papers (aligned with corresponding protocol).

Method	Yaw	Pitch	Roll	MAE
View manifolds [60]	–	–	–	17.52
Random Forests [69]	–	–	–	12.26
Pata. and Cang.* [52]	11.04	7.15	4.4	7.53
MLP + Locations (Ours)	9.56	6.64	4.68	6.96
CNN + Heatmaps (Ours)	**6.19**	**5.58**	**3.76**	**5.18**

Table 3.2 Results on AFLW dataset with 5-fold cross validation. *: Constrains the angles to a certain range.

Method	Yaw	Pitch	Roll	MAE
Kepler [12]	6.45	7.05	5.85	6.45
Ruiz et al. [56]	6.26	5.89	3.82	5.324
MLP + Locations (Ours)	6.02	5.84	3.56	5.14
CNN + Heatmaps (Ours)	**5.22**	**4.43**	**2.53**	**4.06**

Table 3.3 Results on AFLW using testing protocol in [12].

The results clearly show that our CNN-based framework achieves the lowest MAE, significantly improving on the previous state-of-the-art on both the protocols. Interestingly, the MLP based approach also gives competitive performance as compared to previous work. We believe that the exact locations of the facial keypoints, as used in case of MLP, makes it prone to overfitting while the heatmaps act as a regularizer in that sense, giving an edge to CNN based framework. Overall, the experiments provide a strong empirical evidence towards the hypothesis pursued .

3.4 Synthetically generated dataset

Most of the deep learning approaches in head pose estimation rely heavily on large well annotated data. Producing a large and precisely annotated dataset for the task of head pose estimation with the manual annotation is quite difficult. Annotation cost and label accuracy makes manual annotation of real-world head images for creation of large head pose estimation dataset not a good course of action. There have been other approaches for labeling the data more precisely which include method like depth images [13], or inertial measurement unit (IMU) sensors [6]. But both of these methods are susceptible to sensor noise. The BIWI dataset[13], has been used most widely for head pose estimation, has an average error of 1 degree [20]. Another issue with the existing real-world head images dataset for HPE is uneven distribution of samples across different poses. Uniform distribution of these annotated samples is also important for training regression. AFLW does provide us samples of large variety in appearance(e.g., pose, expression, ethnicity, age, gender) as well as general imaging and environmental conditions helping to train our model robustly, but it doesn't provide a uniform distribution of the

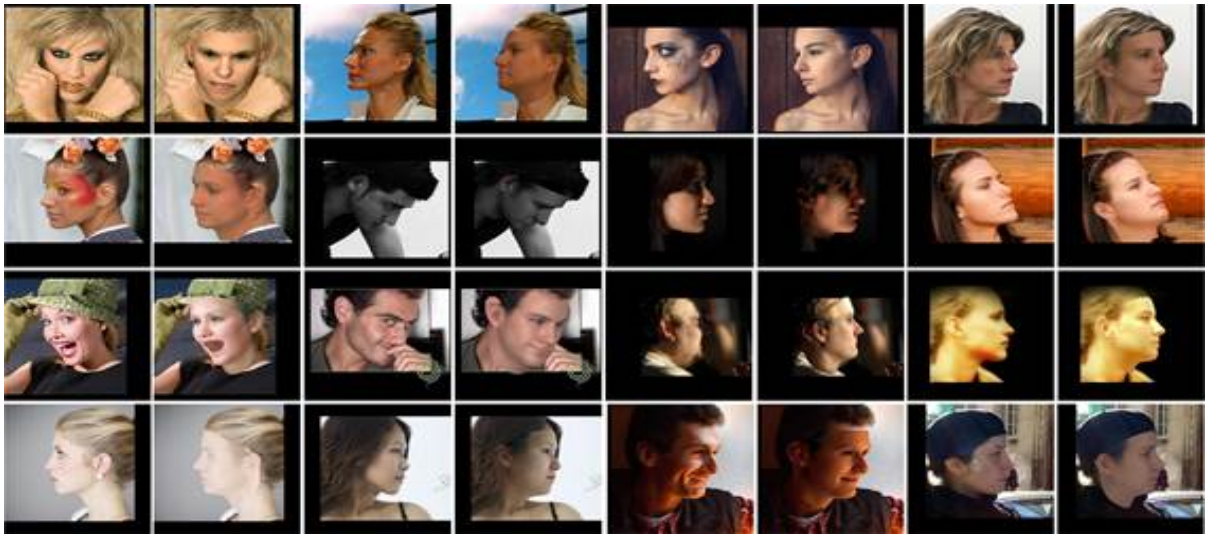

Figure 3.7 Samples from Aflw dataset

annotated samples which is an essential attribute to train our model for HPE. The distribution for the dataset of samples across different angles is shown in 3.9. The samples for BIWI dataset are captured in a controlled environment and relatively are uniformly distributed in comparison to AFLW but isn't uniform enough as amount of samples in extreme poses is far less than amount of samples in neutral pose making it difficult to train more robustly.

Synthetic face images obtained using rendering of 3D face model allows to create a dataset in a inexpensive way and virtually unlimited quantities which are accurately labeled. Moreover synthetic dataset allows us to create a more evenly distributed dataset along all the different poses helping to train the regression model more robustly.

We observe that our intermediate representation of the facial keypoints as uncertainty maps is independent of most of the facial identity features. Synthetically attained face image, in its intermediate representation is indifferent to real face image(with relatively same face structure) for the same head pose. This indifference in intermediate representation is quite helpful for cross dataset training and testing between synthetic dataset and real(AFLW and BIWI) dataset. We use this observation to create a synthetic dataset of our own and use this synthetic dataset to train our model.

3.4.1 Generation of synthetic samples

We choose 10 different 3d head models for generation of our synthetic dataset. The 3D head models used have varied textures and geometrical structures. To render RGB images from each head model we use Unity 3D engine. We render from a perspective view to get realistic head images. The head model is rotated at different angles that follow a uniform distribution. The angles are used as the pose ground truth of synthesized head images. For each 3D head model we create 3600 samples, thus we are able to generate 36000 samples to create a synthetic dataset for head pose estimation.

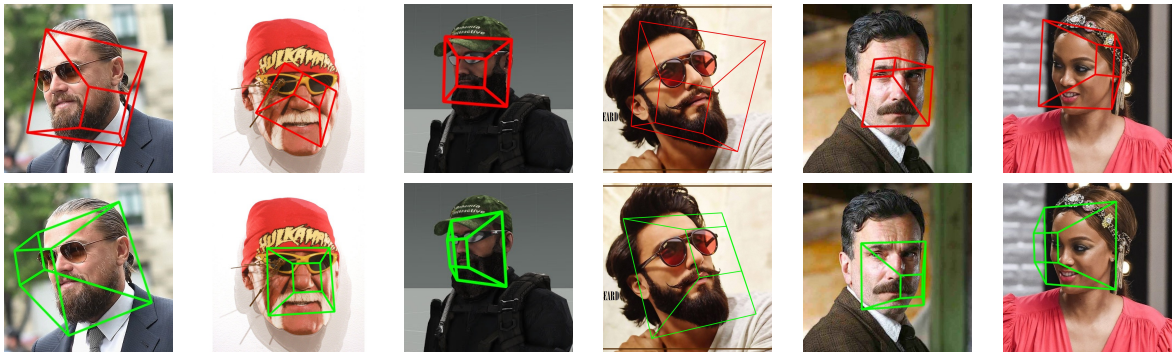

Figure 3.8 Estimation of head pose using three different models (all trained on BIWI), on unseen images taken from the web. **Top row**: Results for CNN-based model [82] which takes RGB images as input, **Bottom row**: Results for our CNN-based framework which takes heatmaps of five facial keypoints locations as input.

Method	Train set	Test set	Yaw	Pitch	Roll	MAE
Ruiz et al. [56]	Biwi	Biwi	6.26	5.89	3.82	5.324
DMLIR [32]	Biwi	Biwi	3.12	4.68	3.07	3.62
CNN + Heatmaps(Ours)	Biwi	Biwi	3.46	3.49	2.74	3.23
CNN + Heatmaps(Ours)	**Synthetic(Ours)**	**Biwi**	**4.23**	**4.64**	**4.06**	**4.31**

Table 3.4 Results with our method using model trained exclusively with our generated synthetic head pose dataset

3.4.2 Experiments and results

We use the same CNN architecture as described in the earlier section. We train our model using our synthetically generated dataset. To test our model for real world images, we use BIWI Kinect Headpose Dataset [13]. We use the entire BIWI dataset(without any split) as test set in out first experiment. The results of this experiment are presented in Table 3.4.

In our second experiment we try to test our model on synthetic images, for this we use Nvidia Synthetic Head Pose dataset[19]. The results of this experiment are presented in Table 3.5.

Though our model fall short in achieving state of the art performance, it does validate our hypothesis that keypoints probabilistic heatmap based learning for head pose estimation is independent of facial features.

Method	Setting	Yaw	Pitch	Roll	MAE
RNN [12]	sequence based	1.78	1.55	1.66	1.67
CNN + Heatmaps (Ours)	image based	**2.41**	**2.56**	**1.94**	**2.31**

Table 3.5 Results on Nvidia Synthetic Head Pose dataset[19]

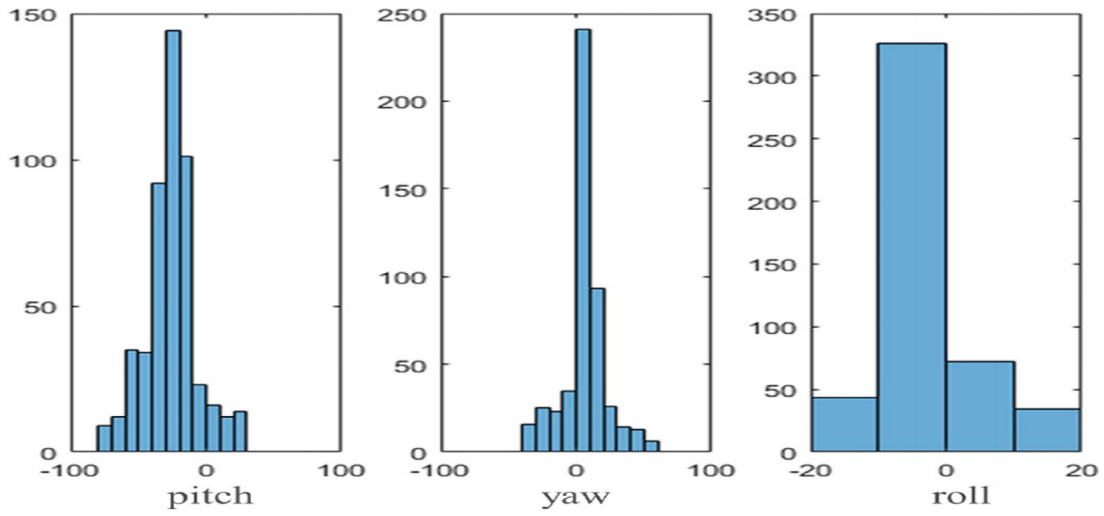

Figure 3.9 Uneven distribution in AFLW dataset(consists of real-world head images) across each angle value, respectively, i.e., pitch, yaw, and roll

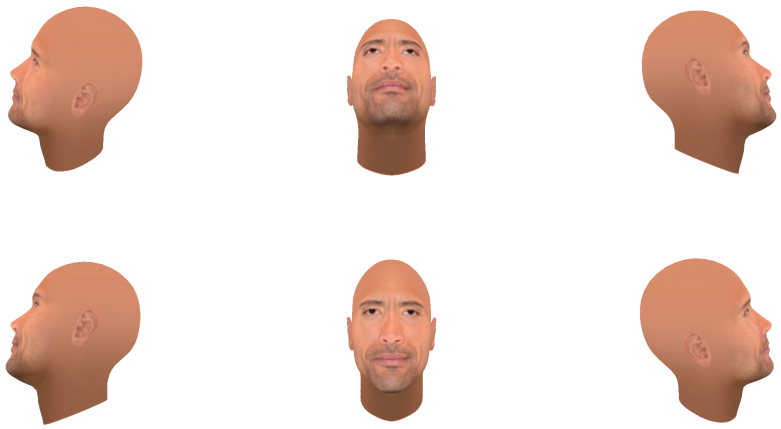

Figure 3.10 One of the 3d face model and some of the samples generated from it by rotation

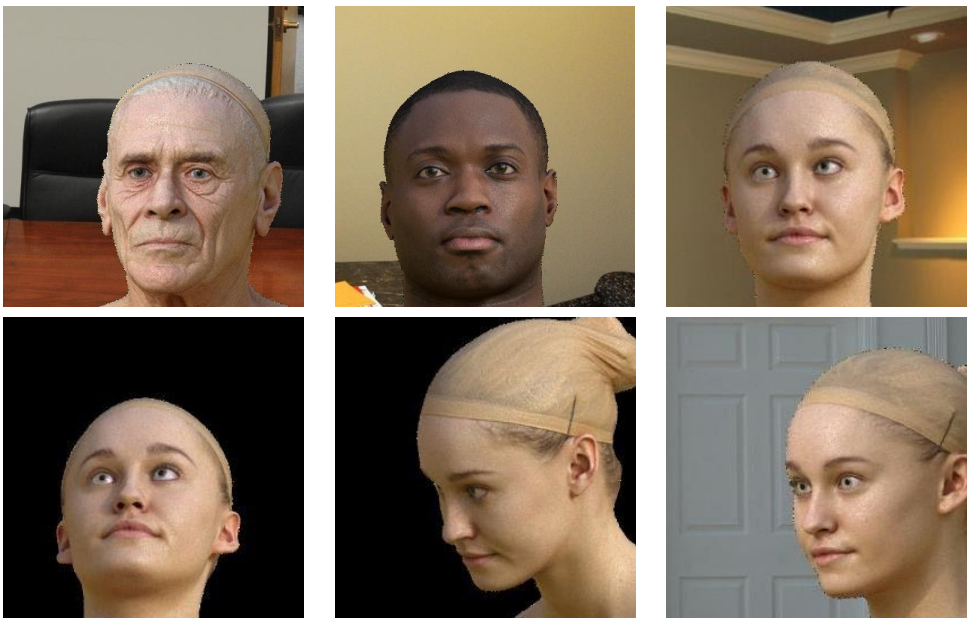

Figure 3.11 Samples from Nvidia Synthetic Head Pose dataset[19].**Top row**: Different face models, **Bottom row**: Head poses for top right face model

Chapter 4

Human emotion recognition

4.1 Facial expression recognition using action units heatmap

Facial expression is one of the most important visual cue for analyzing the underlying human emotions. Despite the continuous research efforts, accurate facial expression recognition under uncontrolled environment still remains a significant challenge. Many early facial recognition datasets were collected under "lab controlled" environments where subjects were asked to artificially generate certain expressions. Such deliberate behavior often results in different visual appearances, audio profiles as well as timing, and is therefore by no means a good representation of natural facial expressions. On the other hand, recognizing facial expressions in the wild can be considerably more difficult due to the visually varying and sometimes even ambiguous nature of the problem. Other adverse factors may include poor illumination, low resolution, blur, occlusion, as well as cultural/age differences. The conventional FER approach is composed of three major steps, i.e., image preprocessing, feature extraction, and expression classification. Such methods based on manual feature extraction are less dependent on data and hardware, which have advantages in small data sample analysis. Deep learning-based FER approaches greatly reduce the dependence on feature extraction by employing an end-to-end learning directly from input data to classification result.

Facial action units (AUs) have been used for facial expression recognition. Action units (AUs) are the fundamental actions of individual muscles or groups of muscles. Intensities of AUs have been described by appending letters A–E (for minimal-maximal intensity) to the action unit number (e.g. AU 1A is the weakest trace of AU 1 and AU 1E is the maximum intensity possible for the individual person). These discreet representation of action units is usually used to predict facial expression. Human emotions lack distinguishable boundaries making it difficult to precisely predict human emotion using action units discreet representation. Moreover these discreet representations are not best for training a model to learn the facial expression using them. We use our experience from head pose estimation where we use probabilistic heatmaps of keypoints. We in our attempt use action unit probabilistic heatmap to train our model. We try to jointly estimate all AU intensities through heatmap regression, along with the location in the face where they cause visible changes. It helps in providing alternative for discreet representation

of action units by having a continuous field and also allow us to incorporate learning of spatial relation between action units.

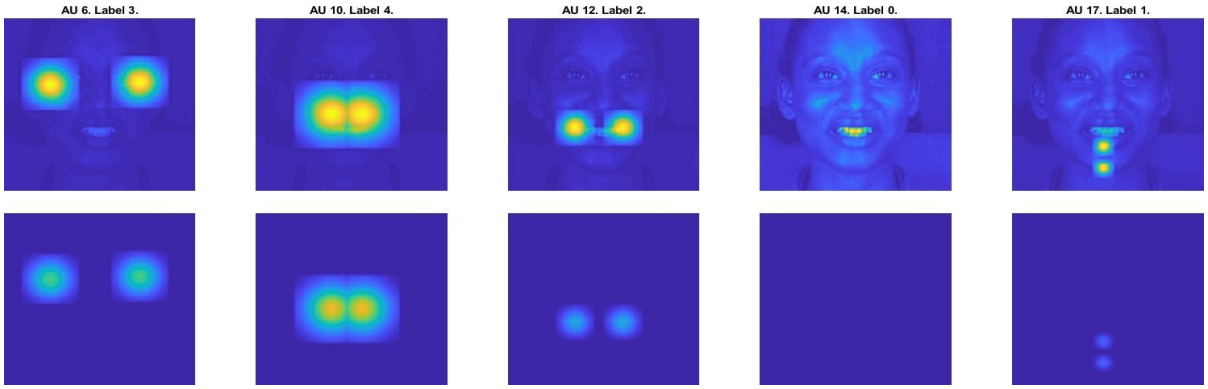

Figure 4.1 Target heatmaps for a particular representative image. The peak and size of the heatmaps are specified by the corresponding labels, and are placed based on landmarks which help in locating Action Unit. These heatmaps are concatenated to construct the heatmap regression.

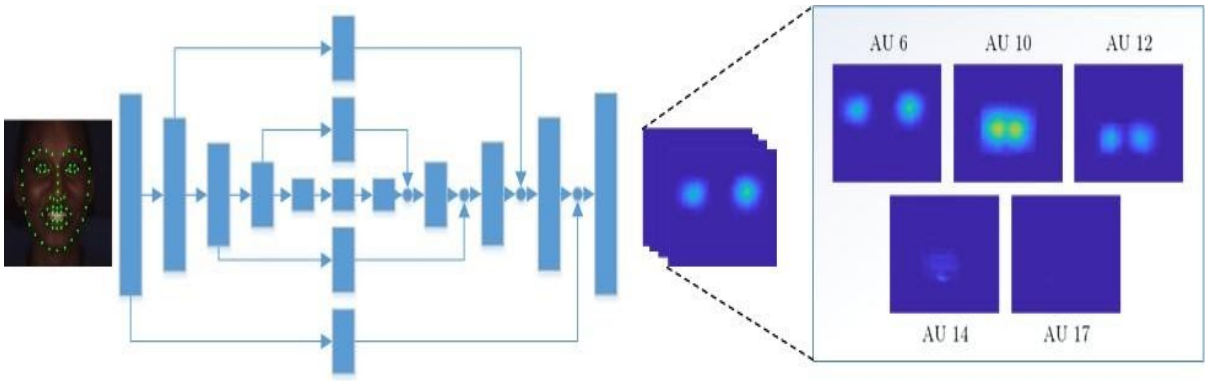

Figure 4.2 The image is processed through the Hourglass network, generating the heatmaps corresponding to the Action Units, at the location they occupy. The heatmaps are activated based on the predicted intensity.

4.1.1 Facial action units heatmaps generation

The facial action units encode the basic movements of individual or groups of muscles that are typically observed when a facial expression produces a particular emotion. The facial emotion recognition system classifies expression categories by inspecting the combinations of the detected face AUs.

We use our experience from head pose estimation where we use uncertainty maps instead of face images to train our model. We use a similar approach here using method in [62] to compute uncertainty maps for action units from face images.

The method jointly estimate all AU intensities through heatmap regression, along with the location in the face where they cause visible changes. It aims to find out a pixel-wise regression function returning a score per AU, which indicates an AU intensity at a given spatial location. Heatmap regression then generates a picture , or channel, per AU, during which each pixel indicates the corresponding AU intensity. to get the ground-truth heatmaps for a target AU, the facial landmarks are first estimated, and a 2D Gaussian is drawn round the points where the AU is understood to cause changes. The amplitude and size of the Gaussian is decided by the intensity of the AU. The utilization of heatmap regression allows learning of a shared representation between AUs without the necessity to believe latent representations, as these are implicitly learned from the info .

In order to try to do so, instead of learning a model to return an output vector with the AU intensities, we propose to use heatmap regression, where different maps activate consistent with the situation and therefore the intensity of a given AU. especially , during training, a 2D Gaussian is drawn round the locations where the AUs are known to cause changes. The intensity and size of the Gaussians is given by the ground-truth intensity labels. the utilization of variable-sized Gaussians allows the network to specialise in the corresponding intensities, because it is understood that higher intensity levels of expressions generally entail a broader appearance variation with reference to the neutral appearance. An example of the heatmaps is depicted in Figure 4.1. Thus use the Hourglass architecture presented in [50], within the way described in Figure 4.2 the heatmap regression architecture, employing a single Hourglass.

4.1.2 Method

CNN with face image as input have been widely used for facial emotion recognition. Our method consists of two parts, first generating the facial action units from RGB face image (described in above section) and secondly, using those facial action units along with RGB image to train our CNN model. Thus the input the input to our model is a 8 channel maps, 5 corresponding to the action unit heatmaps and 3 corresponding to RGB face image.

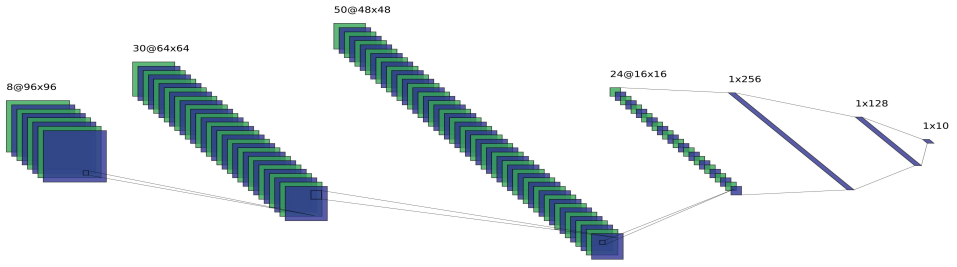

Figure 4.3 CNN architecture used for the facial emotion recognition. The 8 channel input conists of 3 RGB channels and 5 AU heatmap channels

The architecture of our network is described in figure 4.3.

4.1.3 Experiments and results

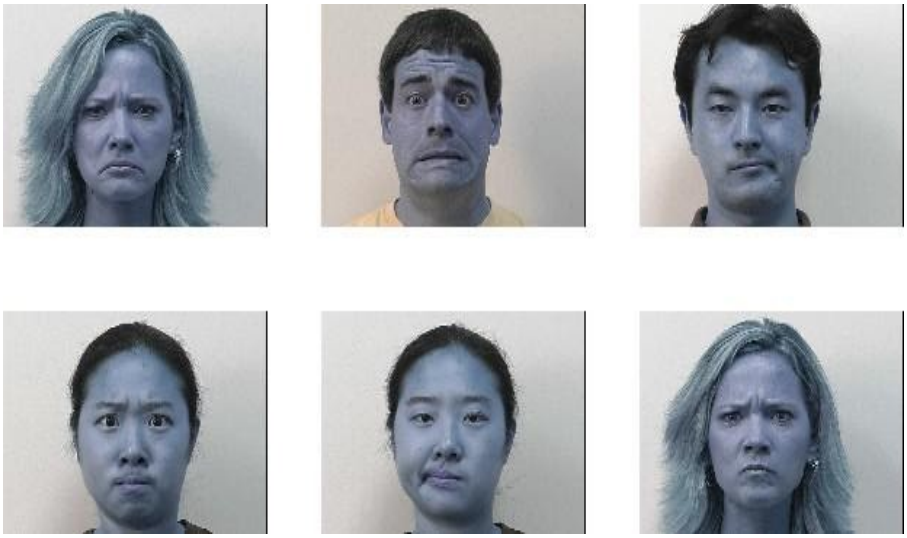

Figure 4.4 Six sample images from CK+ database

Figure 4.5 Four sample images from JAFFE database with different emotion on the same subject

We use two benchmark datasets to measure the performance of our models and test them.

The **CK+** Extended Cohn-Kanade database [39] widely used for evaluating facial expression recognition. The dataset contains 327 labelled sequences collected from 118 distinct subjects. The video sequences are labelled in 7 different emotion which are happiness, sadness, fear, anger, contempt, disgust and surprise. Each sequence consists of 7 frames with expression varying from neutral to peak expression. We take only the last three images as labeled for the expression.

JAFFE [28] dataset consists of 213 face images posed by 10 female subjects of Japanese descent. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects.

Method	Accuracy
Inception [49]	93.2
IACNN [44]	95.37
AU heatmaps (Ours)	**95.84**
DTAGN [27]	97.25
PPDN [92]	97.30

Table 4.1 Average accuracy on the CK+ database for seven expressions classification with 6-fold cross validation

Method	Accuracy
Salient Facial Patch [22]	91.8
AU heatmaps (Ours)	**92.24**
CNN+SVM [59]	95.31

Table 4.2 Average accuracy on JAFFE with 6-fold cross validation

We use these facial action units uncertainty maps to train our model for facial emotion recognition. We use a CNN architecture with 3 convolution layers and 2 fully connected layers. Training is run for 360 epochs with Adam optimizer and set learning rate of 0.00001. We set the batch size to 16. All the experiments are run on a single Nvidia GTX 1080Ti GPU.

We are able to achieve close to state of the art performance using this model on FER datasets CK+ and Jaffe with much simpler and smaller CNN architecture.

4.2 Automated spectrogram annotation for emotion recognition

We identify one of the most challenging problem in human emotion recognition is a large well labelled dataset due to lack of temporal boundaries and different individuals express emotions in different ways. We observe that human expresses their emotion in different form of signals and we can use strong indicator from one of the signal to annotate another.

Humans express their emotion using different cues. We keep our focus on the three forms of expression which are facial expression, speech(tone of voice) and text(used in communication). As mentioned in our previous section, substantial progress has been made to identify human emotion using facial expression. With the application of Convolutional Neural network in learning model for annotated datasets. Identifying human emotion by speech is a tough yet intriguing task. But most of the datasets available are manually annotated thus limiting the amount of samples which can be used for training the model. Another problem with these datasets is that most of them are generated in a constrained environment. We believe that any constrained limit amounts of variation as well as noise thus limiting the performance on new varied samples. Thus we set out to generate a dataset that incorporates vari-

ation, noise and is not limited by manual annotation cost. We observed that movie scenes have these variations as well as the noise we are looking for in our samples. Thus we use different movies to create our dataset.

4.2.1 Proposed approach

Our method consists of two parts. In the first part we generate annotated spectrogram from feature length movies in an automated way. Then we use these annotated spectrogram to train our model.

4.2.1.1 Automated spectrogram annotation

We try to annotate clips from feature length movies based on pre-trained models of facial expression recognition and text-based emotion recognition. To do so we create a pipeline of four sub steps.

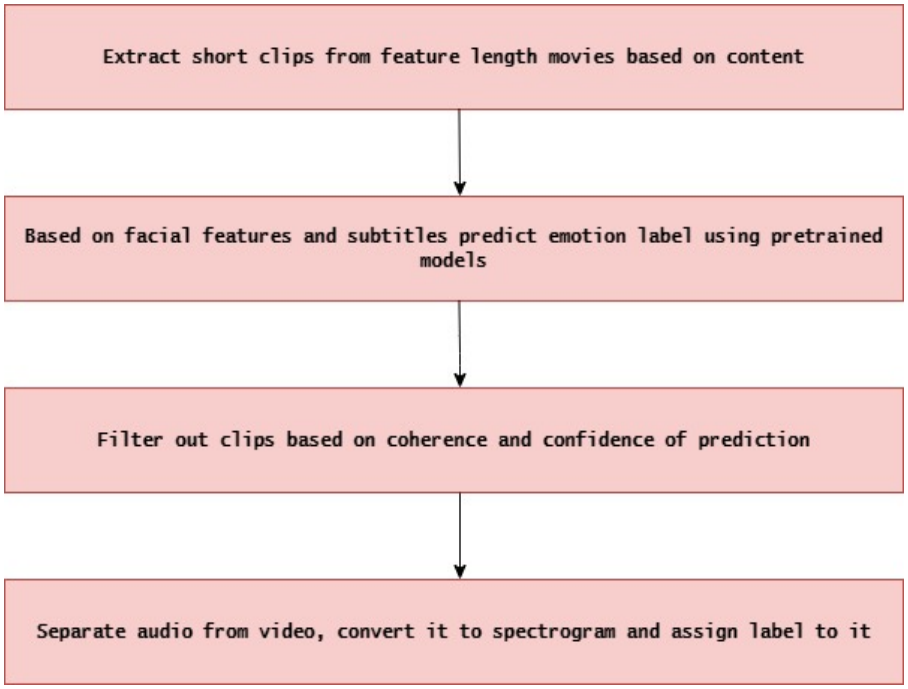

Figure 4.6 Steps in our pipeline for the process of automated annotation of video clips based on emotion

In the first step we cut out clips from the movies based on the content. In the second step we predict based on relevant features from the clips using pre-trained models. In the third step we filter out clips based on the confidence of the prediction. In fourth step we extract out the speech from the clip and convert it to spectrogram. We also attach the label obtained to the spectrogram in this step.

We constrain our clips to have a duration of 3 to 5 seconds(based on length of sentence), consisting of dialogues happening in a same shot. To get clips from the feature length movies we use ffmpeg library. For labelling the clip we use both face images and text in the subtitles. To predict emotion using

facial expression, we first detect all the faces in all the frames for the clip. We then choose only those clips where there is lip movement in between frames so that we consider only the clips where person is speaking. Then we use our model pre-trained on Affectnet dataset to predict facial expression for the detected face images. We then filter out the clips where the confidence of prediction is low as well as the ones lacking temporal context. We then extract out the subtitles during the duration of that clips and use them to predict emotion based on a pre-trained model. If both the facial and textual emotion prediction are coherent we select that clip into our dataset.

From the selected clips, we separate the audio file and convert it to spectrogram(visual representation of the spectrum of frequencies of a signal as it varies with time). Spectrogram has proved to be very successful in training models for speech tasks.

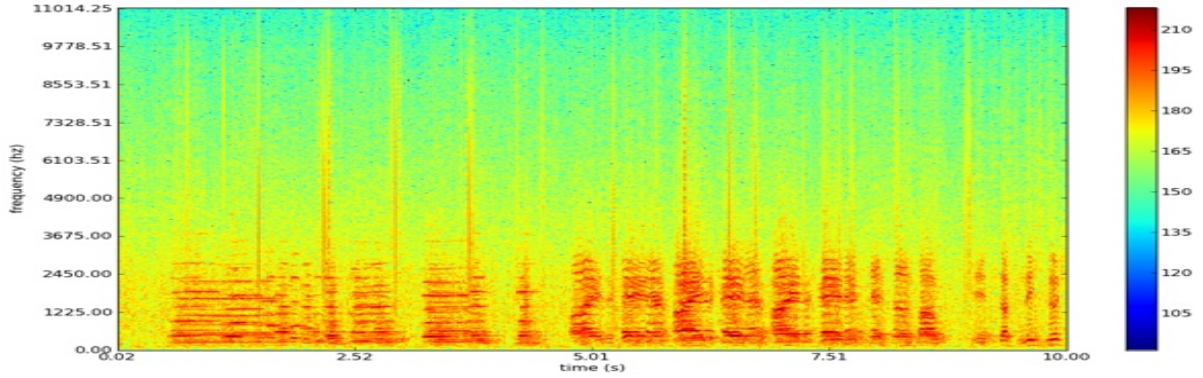

Figure 4.7 Example spectrogram

4.2.1.2 Training using spectrogram

After converting the audio of the clip to spectrogram and obtaining labels, our problem reduces to image classification problem. We try to train these spectrogram for the labels obtained through the automated annotating process described above. CNNs are quite successful in image classification problem and we use them to train our model.

4.2.2 Experiments and results

We use a collection of 20 feature length movies to create our dataset. Our dataset consists of 254 samples annotated in five different emotions(anger, disgust, fear, joy, sadness). The distribution of our samples is shown in table 4.3

We use a CNN architecture with 4 convolution layers and 3 fully connected layers. Training is run for 540 epochs with Adam optimizer and set learning rate of 0.00001. We set the batch size to 16. All the experiments are run on a single Nvidia GTX 1080Ti GPU. We are able to attain average accuracy of 52.3 percent on our dataset with 5-fold cross validation.

Emotion	Number of samples
Joy	84
Sadness	61
Anger	48
Fear	39
disgust	22

Table 4.3 Distribution of samples in our database

The task of applying label to human emotion is quite a complex one. It's not uncommon for human to experience multiple emotions at once and expressing them quite differently in a inconsistent manner. We observe that applying single discrete emotion to a clip is quit difficult. In future, applying multiple continuous label might be more effective way for generating the dataset labelled in automated manner.

Chapter 5

Conclusions

Analysis of human subject in scenes is a complex task. Humans interaction to their environment involves several elements. These interaction vary greatly according to human personality, intent, mood and many other factors. To understand these interactions involves several challenging problems. In this thesis we restrict ourselves to a couple of problems namely head pose estimation and emotion recognition.

In chapter 3 we presented a hypothesis that using an intermediate representation for training our model can give better performance then the conventional way of training the model with face images. For the task of head pose estimation, we are able to accomplish state of the art result using our method. Our method also performs really well with the use of synthetically generated dataset for training. The appreciable performance across cross data training and testing(training with synthetic and testing on real dataset) indicates our method's ability to exclude facial identity specific features and learn head pose relevant features with increased robustness.

In chapter 4 we extend our idea of training our model on intermediate representation for the problem of facial emotion recognition. In this approach instead of using keypoints as we had done for head pose estimation we use facial action units heatmaps as the intermediate representation. We are able to achieve close to state of the art results using this method. Further, we identify scarcity of well labelled large datasets in emotion recognition due to manual annotation and lack of temporal boundaries across different emotions. We utilize the fact that human tends to express their emotion in different form of signals, we use the facial expression and words for annotating the speech signal for emotion. We create a unique pipeline to generate annotated spectrogram from feature length movie videos and create a dataset using this pipeline.

Related Publications

1. **Nose, Eyes and Ears: Head Pose Estimation by Locating Facial Keypoints** [21]
   IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019
   Aryaman Gupta, Kalpit Thakkar, Vineet Gandhi, P J Narayanan
   CVIT, KCIS, International Institute of Information Technology, Hyderabad

Bibliography

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] J. A. Bagnell and D. M. Bradley. Learning in modular systems. 2010.

[3] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2:568–573 vol. 2, 2005.

[4] Y. Bengio, P. Y. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5 2:157–66, 1994.

[5] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5562–5570, 2016.

[6] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara. Poseidon: Face-from-depth for driver pose estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5494–5503, 2017.

[7] P. G. Z. Z. C.-Y. Lee, S. Xie and Z. Tu. Deeplysupervised nets. *AISTATS*, 2015.

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[9] J. Chen, Z. Chen, Z. Chi, and H. Fu. Facial expression recognition based on facial components detection and hog features. 2014.

[10] W. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013.

[11] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis. Robust Head-Pose Estimation Based on Partially-Latent Mixture of Linear Regressions. *TIP*, 2017.

[12] K. et al. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *FG*, May 2017.

[13] G. Fanelli, T. Weise, J. Gall, and L. V. Gool. Real time head pose estimation from consumer depth cameras. In *DAGM*, 2011.

[14] A. Freitas-Magalhães. Facial expression of emotion. In V. Ramachandran, editor, *Encyclopedia of Human Behavior (Second Edition)*, pages 173–183. Academic Press, San Diego, second edition edition, 2012.

[15] C. Gan, B. Gong, K. Liu, H. Su, and L. J. Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5597, 2018.

[16] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[18] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial structures. In *FG NET WORKSHOP ON VISUAL OBSERVATION OF DEICTIC GESTURES*, 2004.

[19] J. Gu, X. Yang, S. De Mello, and J. Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1531–1540, 2017.

[20] J. Gu, X. Yang, S. De Mello, and J. Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1531–1540, 2017.

[21] A. Gupta, K. Thakkar, V. Gandhi, and P. J. Narayanan. Nose, eyes and ears: Head pose estimation by locating facial keypoints. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1977–1981, 2019.

[22] S. L. Happy and A. Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE Transactions on Affective Computing*, 6(1):1–12, 2015.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[24] S. Hochreiter and Y. Bengio. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. 2001.

[25] N. Hu, W. Huang, and S. Ranganath. Head pose estimation by non-linear embedding and mapping. In *ICIP*, 2005.

[26] J. Huang, X. Shao, and H. Wechsler. Face pose discrimination using support vector machines (svm). In *ICPR*, 1998.

[27] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2983–2991, 2015.

[28] M. Kamachi, M. J. Lyons, and J. Gyoba. The japanese female facial expression (jaffe) database. 1998.

[29] P. Khorrami, T. L. Paine, and T. S. Huang. Do deep neural networks learn facial action units when doing expression recognition? *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 19–27, 2015.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[31] J. N. S. Kwong and S. Gong. Composite support vector machines for detection of faces across views and pose estimation. *Image Vision Computing*, 2002.

[32] S. Lathuilière, R. Juge, P. Mesejo, R. Muñoz-Salinas, and R. Horaud. Deep mixture of linear inverse regressions applied to head-pose estimation. In *CVPR*, 2017.

[33] W. Li, F. Abtahi, and Z. Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6766–6775, 2017.

[34] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 103–110, 2017.

[35] W. Lim, D. Jang, and T. Lee. Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4, 2016.

[36] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei. 3d head pose estimation with convolutional neural network trained on synthetic images. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1289–1293, 2016.

[37] X. Liu, W. Liang, Y. Wang, S. Li, and M. Pei. 3d head pose estimation with convolutional neural network trained on synthetic images. In *ICIP*, 2016.

[38] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.

[39] P. Lucey, J. F. Cohn, T. Kanade, J. M. Saragih, Z. Ambadar, and I. A. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

[40] P. M. R. Martin Koestinger, Paul Wohlhart and H. Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[41] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 10(3):325–347, 2019.

[42] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[43] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong. Identity-aware convolutional neural network for facial expression recognition. *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 558–565, 2017.

[44] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong. Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 558–565, 2017.

[45] S. Minaee, A. Abdolrashidi, and Y. Wang. An experimental study of deep convolutional features for iris recognition. *2016 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–6, 2016.

[46] S. Minaee, I. Bouazizi, P. Kolan, and H. Najafzadeh. Ad-net: Audio-visual convolutional neural network for advertisement detection in videos. *ArXiv*, abs/1806.08612, 2018.

[47] S. Mirsamadi, E. Barsoum, and C. Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231, 2017.

[48] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016.

[49] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016.

[50] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[51] M. Osadchy, Y. L. Cun, and M. L. Miller. Synergistic face detection and pose estimation with energy-based models. *JMLR*, 2007.

[52] M. Patacchiola and A. Cangelosi. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, 2017.

[53] G. Peng and S. Wang. Weakly supervised facial action unit recognition through adversarial training. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2188–2196, 2018.

[54] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4929–4937, 2016.

[55] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 439–448, 2016.

[56] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. *CoRR*, 2017.

[57] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vis. Comput.*, 27:803–816, 2009.

[58] Z. Shao, Z. Liu, J. Cai, and L. Ma. Deep adaptive attention for joint facial action unit detection and face alignment. *Lecture Notes in Computer Science*, page 725–740, 2018.

[59] Y. Shima and Y. Omori. Image augmentation for classifying facial expression images by using deep neural network pre-trained with object image database. In *Proceedings of the 3rd International Conference on*

Robotics, Control and Automation, ICRCA '18, page 140–146, New York, NY, USA, 2018. Association for Computing Machinery.

[60] K. Sundararajan and D. L. Woodard. Head pose estimation in the wild using approximate view manifolds. In *CVPRW*, pages 50–58, June 2015.

[61] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[62] E. Sánchez-Lozano, G. Tzimiropoulos, and M. Valstar. Joint action unit localisation and intensity estimation through heatmap regression. 05 2018.

[63] S. Taheri, Q. Qiu, and R. Chellappa. Structure-preserving sparse decomposition for facial expression analysis. *IEEE Transactions on Image Processing*, 23(8):3590–3603, 2014.

[64] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 648–656, 2015.

[65] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *ArXiv*, abs/1406.2984, 2014.

[66] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014.

[67] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017.

[68] M. H. J. B. V. Ramakrishna, D. Munoz and Y. Sheikh. Pose machines: Articulated pose estimation via inference machiness. *ECCV*, 2014.

[69] R. Valle, J. M. Buenaposada, A. Valdés, and L. Baumela. Head-pose estimation in-the-wild using a random forest. pages 24–33, Cham, 2016. Springer International Publishing.

[70] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 149–149, 2006.

[71] J.-G. Wang and E. Sung. Em enhancement of 3d head pose estimated by point at infinity. *Image Vision Comput.*, pages 1864–1874, 2007.

[72] Z. Wang, Y. Li, S. Wang, and Q. Ji. Capturing global semantic relationships for facial action unit recognition. In *2013 IEEE International Conference on Computer Vision*, pages 3304–3311, 2013.

[73] T. Ward and R. Bernier. *Face Perception*, pages 1215–1222. Springer New York, New York, NY, 2013.

[74] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[75] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[76] J. Whitehill and C. W. Omlin. Haar features for facs au recognition. *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 5 pp.–101, 2006.

[77] O. Wiles, A. Koepke, and A. Zisserman. Self-supervised learning of a facial attribute embedding from video. In *BMVC*, 2018.

[78] H. R. Wilson, F. Wilkinson, L.-M. Lin, and M. Castillo. Perception of head orientation. *Vision Research*, 2000.

[79] J. Wu and M. M. Trivedi. A two-stage head pose estimation framework and evaluation. *Pattern Recogn.*, 2008.

[80] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. In *ECCV*, 2016.

[81] S. Yan, Z. Zhang, Y. Fu, Y. Hu, J. Tu, and T. Huang. Learning a person-independent representation for precise 3d pose estimation. In *Multimodal Technologies for Perception of Humans*, 2008.

[82] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson. Face alignment assisted by head pose estimation. In *BMVC*, 2015.

[83] P. Yang, Q. Liu, and D. N. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007.

[84] J. R. Zadra and G. Clore. Emotion and perception: the role of affective information. *Wiley interdisciplinary reviews. Cognitive science*, 2 6:676–685, 2011.

[85] J. Zeng, W. Chu, F. De la Torre, J. F. Cohn, and Z. Xiong. Confidence preserving machine for facial action unit detection. *IEEE Transactions on Image Processing*, 25(10):4753–4767, 2016.

[86] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2013.

[87] Y. Zhang, W. Dong, B. Hu, and Q. Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2314–2323, 2018.

[88] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007.

[89] K. Zhao, W. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit and holistic expression recognition. *IEEE Transactions on Image Processing*, 25(8):3931–3946, 2016.

[90] K. Zhao, W. Chu, and A. M. Martinez. Learning facial action units from web images with scalable weakly supervised clustering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2090–2099, 2018.

[91] K. Zhao, W.-S. Chu, and H. Zhang. Deep region and multi-label learning for facial action unit detection. pages 3391–3399, 06 2016.

[92] X. Zhao, X. Liang, L. Liu, T. Li, N. Vasconcelos, and S. Yan. Peak-piloted deep network for facial expression recognition. *CoRR*, abs/1607.06997, 2016.

[93] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, 2016.