SEMI-SUPERVISED CLUSTERING OF FACIAL EXPRESSIONS

by

Ahsan Jalal

A Thesis Presented to the Faculty of the
American University of Sharjah
College of Engineering
in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science in
Electrical Engineering

Sharjah, United Arab Emirates

November 2017

Approval Signatures

We, the undersigned, approve the Master's Thesis of Ahsan Jalal.

Thesis Title: Semi-supervised Clustering of Facial Expressions.

Signature	**Date of Signature** (dd/mm/yyyy)
_____	_____
Dr. Usman Tariq Assistant Professor, Department of Electrical Engineering Thesis Advisor	
_____	_____
Dr. Hasan Mir Associate Professor, Department of Electrical Engineering Thesis Committee Member	
_____	_____
Dr. Tamer Shanableh Professor, Department of Computer Science and Engineering Thesis Committee Member	
_____	_____
Dr. Nasser Qaddoumi Head, Department of Electrical Engineering	
_____	_____
Dr. Ghaleb Husseini Associate Dean for Graduate Affairs and Research College of Engineering	
_____	_____
Dr. Richard Schoephoerster Dean, College of Engineering	
_____	_____
Dr. Mohamed El-Tarhuni Vice Provost for Graduate Studies	

Acknowledgements

First and foremost, I would like to thank Almighty Allah for blessing me with the opportunity to do my Masters thesis. I would like to pay my gratitude to my academic advisor Dr. Usman Tariq. It has been an honor to work with him on my Master's thesis. His guidance, support, motivation and expertise in the field of deep learning and pattern classification enabled me to achieve my research objectives. I would like to thank my committee members for their interest in my research work. My thanks also go to the faculty of College of Engineering and the Department of Electrical Engineering for providing me with a platform along with the graduate student assistantship that helped me to achieve this milestone.

Dedication

I dedicate this work to my parents who always supported me and believed in me at every step of my life and also to you as a reader . . .

Abstract

Automated facial expressions recognition (FER) is an important area in computer vision and machine learning due to its eminent role in human-machine interaction. FER is key in building intelligent user interfaces, particularly in smart cities. It is also used to enable social robots to naturally interact with humans. However, FER is not trivial as it may vary significantly within different genders, age groups and occasions. Limited availability of the labeled dataset for expression recognition task is another challenge. Therefore, semi-supervised learning algorithm using triplet-loss based deep convolutional neural network is proposed with the motivation to cluster known and unknown facial expressions under unconstrained environment. Faces are detected and aligned from the image dataset and are then used to train various supervised and unsupervised dimensionality reduction methods. Transformed faces in the new dimensions are used for clustering using K-means and consensus clustering. Dimensionality reduction methods that are employed include, principal component analysis, linear discriminant analysis and learning embeddings with deep convolutional neural networks (CNN). The motivation behind using supervised CNN is their ability to learn non-linear transformations in a highly complex feature space. The best results could be found using embeddings that are learned using deep convolution neural networks with consensus clustering method. The novelty of the proposed work is to cluster facial expressions, which were not present while learning the supervised dimensionality reduction methods. Experimental results on two constrained datasets, Multi-PIE face and MMI face datasets, show that the proposed algorithm does not only produce best clustering results on discrete expressions compared to other linear embeddings, but also clusters expressions with different intensities. The proposed algorithm is also applied on a complete unconstrained YouTube dataset and the clustering of different facial behaviors shows that the proposed work can be generalized to non-standard expressions and can learn expression classes from the datasets themselves.

Search Terms: ***Facial expressions, semi-supervised learning, deep convolutional network, consensus clustering***

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Facial expressions are a visible demonstration of the affective state, intention and personality of a person. They play a fundamental non-verbal commutative role in interpersonal relations [1]. Facial expression recognition has a wide range of applications, including, but not limited to, human behavior interpretation [2], electronic customer relationship management [3], social robots, intelligent automobile systems and entertainment industry [4]. Facial expressions have significance in building future human-computer interfaces (HCI) as the present user interfaces lack affective feedback from the user. These interfaces would have the ability to detect subtleties of and shift in user's affective behavior and ability to initiate conversations/interactions based on this information rather than simply responding to user's commands [5], [6]. HCI systems without affective states have interactions which are frequently perceived as incompetent and socially inept. Human computing paradigm suggests that user-interfaces should be human-centered and built on naturally occurring models of human conversations [7]. Lisetti and Nasoz propose a system by combining physiological signals with facial expressions to recognize user emotion and then modify the animated user interface to mirror the user's emotion [8]. In the same view, Kapoor et al. combine the information from a camera, sensor based chair and skin sensor to detect frustration in order to predict when a user needs help [9]. All these aforementioned systems were the initial steps towards affect-based HCI systems. FER systems can provide a mechanism for detecting scenes from movies and social videos which contain expressions of pain, fear and disgust and could provide a valuable tool for violent-content-based indexing of such visual materials and digital libraries [10].

Two approaches are generally used to represent a face, and these are consequently used as facial features for expression recognition. The first approach is the holistic approach where the whole face is treated as a feature space. In their research, Essa and Pentland used the holistic approach to measure facial deformations using optical flow [11]. Nikunj et al. also used holistic approach to capture variation in facial features in temporal domain based on Eigen-face approach [12]. Otsuka and Ohya computed the 2D Fourier transform coefficients on hidden Markov model and optical flow

based model for expression classification in a holistic way [13]. In the second approach, instead of using the whole face as a feature space, one can isolate and use prominent features, such as lips, eyes, eyebrows, cheeks, etc., or sub-regions of face which play a vital role in making unique expressions. Using reference points, facial expressions can be estimated based on the relative positions of these features. These fiducial points can be obtained either manually [14] or automatically from faces [15]. For instance, Stephen and Norman used Facial Action Coding System (FACS), proposed by Ekman and Friesen [16], to develop a system to perform the actions of American Sign Language (ASL) [17]. Mase used optical flow on manually selected facial regions to estimate the motion of facial muscles [18]. Yacoob and Davis took this technique forward and applied optical flow to track the motion of eyebrows, eyes, nose and mouth to classify six basic expressions [19]. In addition to these illustrations, Barlett et al. combined optical flow with principal component analysis to classify facial expressions from still images [20].

A number of researchers have focused on facial key-points as a feature vector for expression recognition task. However, this feature is not useful in unconstrained environment as key-point detection does not work well in bad lighting and noisy images/videos. Same is the case when considering whole faces for expression recognition only, where information regarding expressions may be missed due to highly complex feature space.

The prime goal of the proposed research work is to cluster facial behavior in order to get representative facial expressions which go beyond the basic emotions, such as sadness, happiness, fear, surprise, disgust and anger. Nonetheless, to achieve that ultimate goal, the current research proposes and studies a system to ensure its ability to cluster facial behavior that has samples that are never seen during training of supervised models. Hence, the methodology to achieve this research objective is to detect and align faces, reduce dimensionality using supervised or unsupervised methods and then cluster facial expressions. During learning the transformations for dimensionality reduction, images from certain expressions are excluded and are then included while clustering using various methods. The results of the proposed research work show that it can cluster unknown expressions, i.e. it can associate a separate class for the unknown

expression. It is conjectured that this can be extended to cluster facial expressions in the wild. Some results on a completely unconstrained YouTube dataset are also shown to justify the clustering strength of the proposed algorithm in the unconstrained environment. For dimensionality reduction, Principal Component Analysis, Linear Discriminant Analysis and learning non-linear embeddings using Deep Convolutional Neural Networks are used. For clustering, K-means and Consensus clustering; drawing the motivation from genetics clustering where classes need to be automatically discovered from the data are used.

1.1. Background

In daily communications, it is an everyday task to recognize facial expressions without any effort as the human mind is well trained to do this job. Humans can easily identify age range, gender and expressions from someone's face and get a response that affects words selection for conversations with people [21]. Facial expressions and other gestures convey non-verbal communication. They also compliment spoken words towards the listener to elicit the intended meaning from the speaker. Therefore, emotions shown in facial expressions play a vital role in daily social life even without the notice. Happiness, sadness, surprise, anger, fear and disgust are known as basic emotions that are communicated via facial expressions. Facial expressions have a significant effect on listening interlocutor. Fifty five percent of speaker's effectiveness is contributed by facial expressions while 38% percent is conduced by voice inflection, and just 7% is influenced by the spoken words [22].

Nonetheless, for computers, it is not an easy job to recognize expressions from face and translate them to specific mood-based conversations. This side of communication is hard to establish between humans and machines. Progress in this field will ensure more effective human-machine interaction in near future [23]. In this research [24] Pentland examines the mathematical tools that have proven to be useful in describing the taxonomy of the problem domain. He highlights the significance of smart user-interfaces. Van and Andy [25] point out that automatic recognition of facial expressions will help to establish natural human-machine interfaces or conversational

interfaces [26]. Similarly, studies [1] and [11] indicate that the automatic classification of facial expressions may help to study behavioral science.

1.2. What are Facial Expressions?

Facial expressions are temporarily deformed facial features, such as eyes, nose, cheeks, eye brows and skin texture, by the contraction and relaxation of muscles and movement of associated bones. They are the result of facial muscle actions which are triggered due to the nerve impulses generated by the brain, based on the basic senses or abstract thoughts. They last for few seconds but rarely more than 5 seconds or less than 250 milliseconds as reported in [27]. They provide a medium to express felt emotions, non-verbal communication and physiological conditions. Studies on facial expressions started in the nineteenth century. In 1872, Darwin proposed the concept of universality of facial expressions and their continuity in man and animals and claimed that there are some particular emotions that originate from associated habits [28]. In 1971, Ekman and Friesen proposed six primary emotions and linked each of them with a unique facial expression. This set of emotions is referred to as *basic emotions* [29]. They are

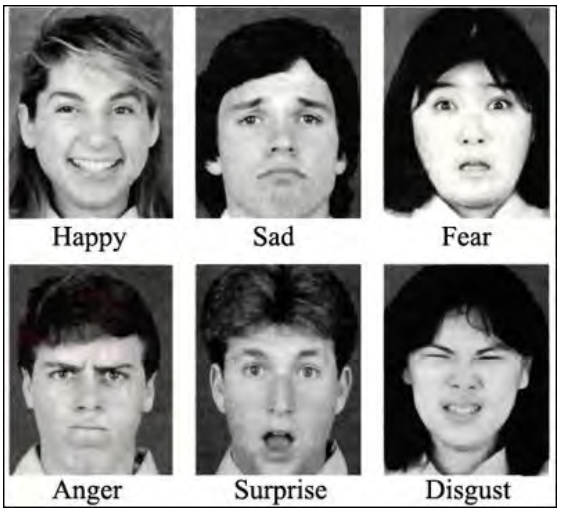

Figure 1: Six basic emotions represented by unique face expressions, adopted from [30]

generally universal across human ethnicities, genders and age groups. The set contains feelings of sadness, happiness, anger, surprise and disgust as shown in Figure 1. Ekman

and Friesen also developed Facial Action Coding System (FACS) to describe facial expressions by human observers. FACS describes thirty two atomic facial muscle actions, named Action Units (AU), and fourteen other additional Action Descriptors (AD). AU are represented in terms of visible appearance changes, so they are the prime candidate for computer vision based expression detection. An example image with associated AU is shown in Figure 2.

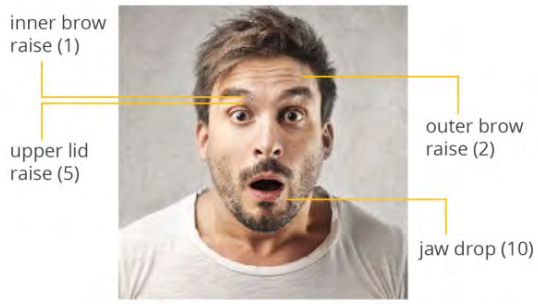

Figure 2: Expressive image with associated AU and their physical interpretation which objectify the facial expression, adopted from [31]

1.3. Automatic Facial Expressions Classification

In the past, FER was the main research field for psychologists until Suwa et al. [32] proposed an initial form of automatic face expressions using image sequence in 1978. Automatic FER gained much inertia two decades ago due to the advancements in face detection, face tracking and recognition systems.

The standard algorithmic pipeline for automatic facial expressions recognition is divided into three steps: pre-processing (which includes face detection, alignment, normalization and face registration), feature extraction and machine learning.

1.3.1. Face detection. Face detection is the first step in an automatic facial expression recognition task. Given a random image, the goal of any face detector is to determine the presence of face(s) in the image and then return the coordinates of the geometry bounding that face as shown in Figure 3. The aim of pre-processing is to align and normalize visual information in such a manner to enhance the semantic meaning

of the feature extracted in the later part. This step eliminates fundamentally irrelevant variations in the input image coming from misalignment to alleviate the effect of head pose variation and identity. Mean and variance normalization is done to reduce the effect of lightness and contrast variations; this is demonstrated in Figure 4. However,

Figure 3: Multiple face detection with bounding boxes

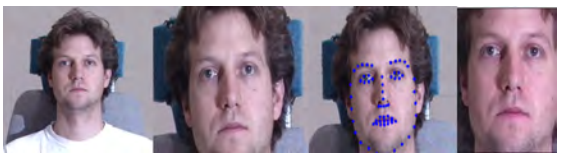

Figure 4: Input image with face is used to extract face, calculate facial landmarks which are used to align, extract and scale face to pre-defined parameters

face detection is not as simple for computer algorithm as it is for humans as the latter are well-trained for this job and can analyze faces effortlessly. It is challenging for a computer algorithm to detect faces under blur, occlusion, scale and variation in illumination or facial features such as closed eyes. Most of the algorithms work on frontal view of face. After face detection, orientation and scale of the test face is determined by comparing it with a model face.

Face detection has undergone significant development after the seminal work of Viola and Jones [33]. Its advantage over other algorithms is its real time detection with high accuracy. It has four stages: HAAR feature selection, creation of integral image, AdaBoost training and cascade classifiers. Some other state-of-the-art face detection

algorithms include [34] and [35] which proposed mixture of deformable part models [36]. New face detectors can easily detect frontal faces and are widely used in digital cameras and social applications, such as Facebook.

1.3.2. Feature extraction. Feature Extraction is the most crucial phase for a FER system as the accuracy of the classification step is primarily dependent on the selection of a good feature set. Facial feature extraction attempts to find most suited representation of faces for recognition. The prime challenge is the nuisance factors such as head pose variations, illumination or even alignment errors that have larger impact on the appearance than the expressive behavior [37]. There are three approaches for feature extraction: comprehensive spatial feature-based template-matching systems (also called appearance features), geometric feature-based systems and motion (hybrid) features.

Various appearance-based features are used in the literature for facial expression recognition task; for example, Local Binary Pattern (LBP) [38], Pyramid of Histogram of Gradients (PHOG) [39], Local Phase Quantization (LPQ) [40], Local Phase Quantization-Three Orthogonal Planes (LPQ-TOP) [41], Local Binary Pattern-Three Orthogonal Planes (LBP-TOP) [42], Scale Invariant Feature Transform (SIFT) [43], Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [44], Graph-preserving sparse nonnegative matrix factorization (GSNMF) [45] and Gabor filters [46]. In appearance-based approach, template can be a whole image or regions of the pixel image, and feature vectors that are obtained after processing on the raw features. Principal component analysis (PCA) [47], stacked auto-encoders and various dimensionality reduction methods can be used for dimensionality reduction of the feature space. More recently, people have focused towards a more data-intensive method to extract holistic features from data itself and jointly learn features and classifiers, i.e., deep learning. Some of the works in this domain are discussed in section 2.3.2.

In geometric approach, face key-points or landmarks are detected in the images as shown in the Figure 5. The distance between the feature points and the relative size of the face components form a feature vector which can be used to form a geometrical representation of the faces. The geometric approach is robust as com-

pared to appearance-based approach in terms of scale, orientation and location of the face; however, challenges could be faced in unconstrained lighting and pose variations. Different geometric based approaches used for facial expression recognition include Piecewise Beizier Volume Deformation (PBVD) [48], Candide Facial Grid [49], Geometric Distance [50], Extended Dynamic Mesh (EDM) [51], Curvature maps [52], optical flow [53], Free-Form Deformations (FFD) [54], Level curve deformations [55] and Basic Facial Shape Component (BFSC) [56].

Motion features are constructed based on a dense registration of appearances between consecutive frames [57]. Several distinct cues are displayed on the face for each specific emotion; for instance, stretching of lips, movement of eyebrows, etc. [58]. Deep Bidirectional Long Short-Term Memory Recurrent Neural Network (DBLSTM-RNN) is a motion based approach used to predict the continuous value of emotions from audio and visual modalities [59]. Researchers have also modeled temporal dynamics from motion features for facial expression recognition. For example, [60] estimates whether a facial behavior is deliberate or spontaneous from temporal features.

Features are selected to extract expression information from the images since they are the base for the classification process. Quality of classification is dependent primarily on the feature set selected for this task.

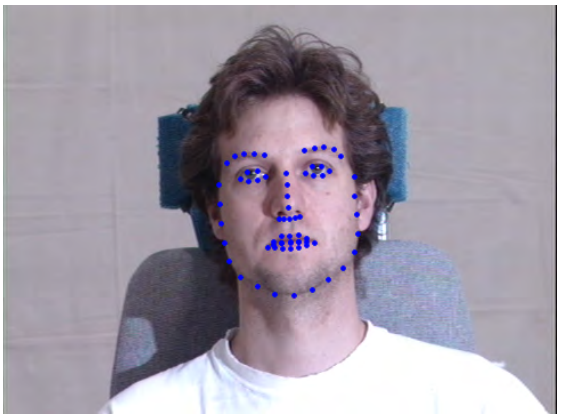

Figure 5: Face Keypoints are used as a feature set for facial expression classification

1.3.3. Machine learning. The last step in automatic FER systems is machine learning, which is performed on the selected features to identify expression in the pro-

vided face image. Learning techniques are broadly divided into three classes: supervised, unsupervised and semi-supervised learning. Supervised learning is a paradigm of deducing a function based on labeled (supervised) training dataset. Each example in supervised learning is a pair that consists of an input object and its ground truth, generally known as supervisory signal. Supervised learning algorithms learn a classifier which can be discrete (classification) or continuous (regression) based on the output. Accuracy/ purity of the classification can be measured given ground truths for the test dataset. Common techniques used in supervised facial expression recognition are Logistic Regression [61], Support Vector Machines (SVMs) [62], Neural Networks [63] and Linear Discriminant Analysis [64]. On the other hand, unsupervised learning is a paradigm of learning hidden structures from the unlabeled dataset as there is no teacher vector (supervisory signal). Semi-supervised learning falls in-between supervised and unsupervised learning techniques. Some techniques used in this domain for facial expression recognition are: Supervised Locally Linear Embedding (SLLE) [65], Semi-Supervised Aligned Cluster Analysis (SSACA) [66], ultra-large scale clustering of temporal event [67], Fuzzy C-Means clustering [68], K-Means based segmentation of faces [69] and Clustering based Discriminant Analysis (CDA) [70].

Zhao et al. proposed Supervised Locally Linear Embedding (SLLE) to classify facial expressions into basic emotions in Independent Component Analysis (ICA) transformed space [65]. They used Japanese Female Facial Expression (JAFFE) database consisting of 213 frontal face images from seven expressions (six basic expressions and one normal expression). Generalized Regression Neural Networks (GRNN) were then used on transformed features to learn mappings which were classified into different expressions using K-NN algorithm. The researchers achieved 88.56% and 89.99% recognition rates in the original space and ICA space, respectively. Vandal et al. [67] used temporal clustering of facial events for eyebrow raiser, eyebrow lowerer and smile. They proposed their own dataset, comprised of more than 1.5 million facial videos. HOG features were extracted from the region of interest (ROI) within the frame from each video, and then SVM classifier with radial basis function (RBF) kernel was applied to compute facial metrics. Finally, K-means clustering was run on eyebrow raiser, eyebrow lowerer and smile features, separately. Eyebrow lowerer events were grouped

into four clusters while the other two events were grouped into five clusters. Smile expression achieved a maximum true positive rate of 86% while eyebrow lowerer and eyebrow raiser got 65% and 62%, respectively. Senthilkumar et al. [69] used K-means based segmentation to extract features for facial expressions recognition. Araujo and Kamel [66] proposed semi-supervised temporal clustering for facial expressions recognition by adding pairwise constraints as a side information to boost clustering process. They used 1,486 annotated facial images from VAM corpus spontaneous facial-emotion database that consisted of three emotion primitives: activation (calm-excited), dominance (weak-strong) and valence(positive-negative). The researchers reported a performance improvement in the range of 0-15 percent across 20 speakers when comparing SSACA to ACA. Active Appearance Model (AAM) based features with Semi-supervised Fuzzy C-Means (FCM) clustering on discrete emotions was proposed by Liliana et al. using limited constrained image dataset [68]. They used relative positions of facial landmarks as shape feature. They used 209 images from CK+ dataset as training images with 8 emotion classes. The researchers tested their proposed approach with 15 images per class. They were able to achieve an average accuracy of 80.71% using fuzzy c-means.

Chen et al. [70] used clustering for feature extraction and classification with nearest neighbor. They used 1,428 images from AR face database from 119 subjects with three facial expressions (neutral, smile and anger). They experimented with binary classification of expressions: neutral versus non-neutral, smile versus non-smile and anger versus non-anger. After learning these classifiers, the researchers combined the results to get neutral versus anger, neutral versus smile and smile versus anger classifiers. They were able to achieve classification rates of 86.7%, 98.2% and 89.1% for neutral versus non-neutral, smile versus non-smile and anger versus non-anger, respectively.

Prediction of expression can be targeted at frame-level or sequence-level. In the former case, separate label prediction is given for every single frame; whereas, one label, possibly multidimensional, is assigned to sequence of frames in the latter approach. Sequence-based problems can be solved by using sequence-based classifiers, such as Hidden Markov Models (HMM) [71] and [72], or a majority vote of a frame-level clas-

sifier can be used as described in [73]. Other techniques; for example, multiple-instance labeling, have been also proposed [74].

Frame-level labeling can be performed using Ensemble learning techniques based on co-association matrices of the data [75], Support Vector Machines (SVMs) (can be linear or non-linear) [62], logistic regression [76], Convolutional Neural Networks (CNN) [77], density regularization [78] and Active Appearance Models (AAM) [79]. Examples of some multi-class classifiers for FER systems that are used in previous research are: AdaBoost [80], Linear Discriminant Analysis (LDA) [81] and multiclass SVMs [[46], [82], [64]]. They are used due to the complexity and types of facial expressions. Alternatively, multiple binary classifiers can be used where multiple classes can be active simultaneously as proposed in [83]. All aforementioned techniques have strong correlation with temporal dimension, so they can be exploited in consecutive frames since an expression usually lasts for more than one frame.

Feature fusion is another interesting dimension in which more than one combination of feature type and representation strategy are considered. Different features are considered in the experimental setup rather than studying which is the best-performing feature. The problem is then defined as finding the best combination of different feature types and their representations [84]. This knowledge of fusion can be extended to find the optimal fusion strategy for two or three dimensional information [85] which can be used as a learning problem as proposed in [86]. There are other unsupervised learning techniques using features such as Principal Component Analysis (PCA) which are generally used due to large dimensionality of feature vector representing the face.

Recently, a great deal of the research work in expression recognition has been shifted towards a data intensive method, known as Deep learning (DL). The interest in DL techniques has enormously increased, particularly when a deep learning algorithm by Krizhevsky et al. [87] achieved state-of-the-art accuracy on the ImageNet dataset. In general, when a hierarchy of features is trained, algorithms are called deep models. Some of the notable works on deep convolutional neural networks (DCNNs) in facial expression recognition include [88], [89], [90], [91], [92] and [93].

The eeaders are referred to [94], [95] and [37] for a comprehensive survey on facial expression recognition.

As pointed out earlier, FER community is primarily divided into two streams: holistic approach and action units based approach. In the holistic approach, the whole face provides a single input to recognition system. The system then classifies the input to discrete facial expressions, such as happiness, sadness, anger, surprise, disgust or fear. In the second approach, automated facial action units or facial key-points are identified for expression recognition purpose. However, to the upmost of the researcher's knowledge, no research work in this field which looks beyond these handful set of expressions or combine these two lines of researches is found during the literature review. Nonetheless, it is important since several facial action units are not independent of each other, i.e., if one region of the face moves, it is possible that another part of the face also move with it. For example, stretching the lips can raise cheek muscles and stretch eye's action units as well. This loss can affect the robustness of automatic HCI systems due to limited information of the user's emotions. Hence, the question to address here is whether it is possible to look beyond the six basic emotion categories and learn clusters of facial behaviors in an unsupervised or semi-supervised way. To answer this question, available techniques leveraged in the current research work to develop a new algorithm for facial behavior categorization under unconstrained environment.

In order to achieve the ultimate goal of clustering unconstrained facial behaviors which are not limited to six universal emotions, it is necessary to prove the ability of the proposed algorithm to cluster those expressions which are never used during training of supervised models. In order to achieve this, faces are detected from images and videos and are aligned for feature extraction. Face detection is based on Histogram of Oriented Gradients (HOG) [39] and linear SVM. Affine transformation is used by geometric features to align faces based on fiducial points of eyes, nose and lips. Raw pixel, LDA, PCA and deep convolutional neural network based features are used for the tests. PCA and LDA are used to learn linear embeddings which effectively reduce dimensionality of data. CNN are used to learn non-linear embeddings of the raw pixel feature space. The learnt embeddings are 128 dimensional. Finally, clustering is performed using K-means and consensus clustering. Consensus clustering is obtained using multiple K-means run on the same data with different cluster sizes to find a single clustering which better separates known and unknown expression classes. During learning deep

CNN based transformations for dimensionality reduction, samples are removed from one of the known classes; they are added later in the test dataset along with other classes (present during training) to perform clustering using aforementioned techniques. Results obtained substantiate the approach of clustering unknown classes into separate clusters. It is proposed that the process could be extended to solve real world problem of clustering unknown expressions in complete unconstrained environment. Some results obtained from the algorithm are applied on a complete unconstrained dataset taken from YouTube under creative content license. Results show that the proposed approach can cluster expressions in the wild, which can be further generalized in future to learn more refined expression classes from the dataset themselves.

1.4. Novelty

The current study fills a gap in the literature. It is different and novel compared to other approaches used for FER because, unlike [67] and [66], static images are used in the experimental setup as the interest is in the differences in appearance across images. Moreover, unlike [68], appearance-based are used instead of key-point tracking. Apart from any other work in semi-supervised clustering of facial expressions, clustering of unknown facial behavior from the dataset is intended. It is conjectured that embeddings learnt on known expressions may also be useful to cluster unknown expressions. This research serves as a proof-of-concept. As the experimental setup is considerably different from other work on FER reported in the literature, research results are expected to be diverse and different from those obtained from research on the topic.

This proposed algorithm is extended to cluster facial behavior in the wild. Moreover, preliminary results on an unconstrained dataset collected from YouTube are also reported to show the clustering of facial behavior which can be used for further analyses like conversations.

1.5. Thesis Organization

This thesis is organized as follows. Chapter two discusses the methodology used for semi-supervised clustering for unknown facial expressions. In chapter three, a review of the databases used in the experiments is given. Experiments and results of semi-supervised clustering techniques are shown in chapter four. Finally, chapter five highlights the main remarks of the current study and suggests some recommendations for possible future works.

Chapter 2: Semi-Supervised learning for facial expressions

In order to select the best feature for a given problem, raw features are primarily the first choice in this regard. Nevertheless, it is not guaranteed to be the best feature in each case due to several reasons such as complexity of the feature space. If required features are naive compared to other features in the data, it is possible that clustering techniques opt for undesirable features and may not give required results. The prime objective of this research is to cluster facial expressions which are not as strong as other features describing gender, skin color, etc. Therefore, a low-dimensional space is needed to bring similar expressions closer to each other and move different expressions far apart.

PCA on raw features is used to reduce dimensionality while keeping information loss minimum by selecting the strongest and data-defining features. Projections in PCA's feature space ensure maximum variance within data is kept, which helps to cluster different classes in the data. Since PCA is an unsupervised algorithm, it selects the strongest features without any information of the objective to extract expression features from the data. Features in raw pixel domain are complex in nature and expression features are mixed with other strong facial features; therefore, clustering can produce undesirable results when performed on PCA components.

In order to bring similar expressions closer to each other, supervised LDA algorithm is used on raw features. The main objective of LDA is to minimize intra-class distances and maximize inter-class distances using training dataset to compute transformation matrix. Projected data from LDA then clustered to get the required expression classes. LDA performs optimal only when data is Gaussian distributed with equal covariance. Moreover, raw features space is considerably complex to separate naive expression features even with supervised technique. Hence, a non-linear dimensionality reduction is needed to address the problem.

Raw dataset is complex in nature, so it requires an algorithm which can learn features non-linearly and extract the complex association between features and the given class. Neural networks are used for their ability to learn complex relations within dataset. These are designed on a biological nervous system and made up of a large

number of interconnected nodes called “neurons”(a specialized cell transmitting nerve impulses) which work in unison to solve problems like humans nervous system do. It is widely used for pattern recognition tasks in computer vision, which motivated the researchers to use ANN to train model on selected expressions. Trained model can be used to produce embeddings/mappings on test dataset in such a way that different expression classes are separated, which will ease clustering onwards. The final objective of the current research is to cluster those expressions that are not in the learning dataset. Thus, triplets of data are used as input to the ANN model that has one matched and one unmatched expression pair. The objective is to bring matched pairs of image class closer to each other and vice versa for the unmatched pairs. Model trained on triplets is more generalized compared to labeled images as it works on matched/unmatched pairs without knowing the actual class. Therefore, it is more general and suits the given problem well.

After feature selection, clustering needs to be performed to distribute data into respective classes. K-means clustering is a very popular clustering technique in cluster analysis and data mining. It distributes data into k classes, where k is specified, in which each observation from the dataset belongs to the cluster with nearest mean value. Sometimes clustering techniques cluster dissimilar objects in same partition due to the formation of complex decision boundaries based on feature space of the input data. Since the total number of expression classes in the test dataset is an unknown parameter, clustering on the dataset cannot be done with 100 % surety. Consensus clustering technique helps in aggregation of partitions with the objective to get a single partition of data with better quality. Similarity matrix is computed for the dataset using the information of each observation position in different partitions. This similarity matrix results into a single partition in which similar objects are closer to each other, which leads to a better clustering with improved purity level compared to simple K-means.

The aforementioned techniques are explained in-depth to show how they extract features and cluster them.

2.1. PCA

PCA is an unsupervised dimensionality reduction algorithm proposed by [96] to reduce n dimensional feature vector x_k to a new reduced feature space, using a projection matrix W. In PCA, transformation matrix W (consisting of orthonormal eigenvectors of the total scatter matrix S_T) is computed such that if the data is projected onto this matrix $z = W^T x$, variance in the projected data is maximized $Var(z) = W^T \sum w$. For instance, for the first principal component, the following objective function will be solved as given in Equation 1.

$$max_{W1}[W_1^T \sum W_1 - \alpha(W_1^T - 1)] \tag{1}$$

Gradient of Equation 1 with respect to W_1 is calculated and simplified to $\sum W_1 = \alpha W_1$, where α and W_1 are the first eigenvalue and eigenvector, respectively. The later eigenvectors are computed by adding an orthogonality constraint to Equation 1. It can be seen that the eigenvectors of the covariance matrix are the solution, i.e., the directions which maximize the variance.

PCA is used to reduce the feature space to 100 components for maximum representation of the input data with less information loss.

2.2. LDA

LDA is a supervised dimensionality reduction algorithm which projects the n-dimensional feature vector x_k to a new reduced feature space so that when data is projected, classes are well separated. Transformation matrix W is used to project data into low-dimensional space where $W \in R^{n \times m}$ containing m eigenvectors v corresponds to m largest eigenvalues of the Equation 2.

$$W = S_w^{-1} \times S_B \tag{2}$$

where S_w is the total within class scatter matrix and is defined as $S_W = \sum_{i=1}^{k} \sum_{t=1}^{n} (x^t - m_i)(x^t - m_i)^T$ and S_B the between class scatter matrix defined as $S_B = \sum_{i=1}^{k} N_i (m^i - m)(m^i - m)^T$. LDA tries to maximize inter-class separation and minimize intra-class

separation for better clustering. Thus, LDA is used on training dataset since it is required to maximize separation between emotion classes irrespective to other strong facial features like individuality, skin tones, gender, ethnicities, etc.

2.3. Non-linear embeddings using deep convolutional neural network

Convolutional neural networks (CNN) have a huge impact on computer vision society as it has improved many state-of-the-art in various applications, such as face recognition [97]. For the problem of clustering unknown facial expressions, is is necessary to have an embedding of the dataset that can bring different expressions farther and similar facial expressions closer to each other even without knowing the actual class labels, which is in spirit of triplet-based deep CNN. This technique differs from its classical deep CNN variant in its use of 'triplet-based 'loss, where a pair of similar facial expressions $(a,\ b)$ and a third dissimilar facial expression (c) are compared. The objective function here is to make (a) closer to (b) than (c). In other words, contrary to other metric learning approaches, comparisons are always relative to pivot expressions. This method currently achieves the best performances on LFW and YTF face datasets [97].

In application of dimensionality reduction techniques, such as PCA, LDA applied directly onto raw pixel features sometimes takes off decisive features for expressions as these are somewhat sensitive and less prominent compared to other facial features like individuality, skin texture and color, gender, regional characteristics, etc. In order extract mappings/embeddings which are expression specific for clustering purpose, triplet-loss training using artificial neural network is used.

2.3.1. Artificial Neural Networks. Artificial Neural Networks (ANNs) and their applications in different fields of life are one of the most researched topics in modern times. For instance, ANN has found its application in object detection and recognition [98], face detection [99], facial expression recognition [100], smart electronic gadgets [101], automatic cars [98], robots [102] and optimization problems [103]. ANNs are inspired by the biological neural networks that are found in animal brains [104]. The reason for their popularity is that, although they are conceptually simple, they can learn considerably complicated non-linear decision boundaries. Hence, they can be applied

for complicated decision making purposes. Another important property that sets them apart from other algorithms is their ability to learn data-driven features, i.e., they are capable of discovering the features that are useful for classification autonomously [105]. Particularly, with the advent of deep learning and deep neural networks (multi-layer neural networks), the AI field has seen massive improvements in a number of benchmark datasets. Thus, there is no apparent reason that methods based on deep neural networks will not improve the current state-of-the-art methods of recognizing facial expressions.

One of the most important early breakthroughs in deep learning came in terms of a successful greedy layer-wise training of Deep Belief Networks [106] using Restricted Boltzman Machines in 2006. Later, such a performance improvement is also shown for Deep Autoencoders [107]. However, one of the most important results in deep learning, since 2012, are achieved by Deep Convolutional Neural Networks (Deep CNNs) [105]. Deep CNNs achieved a remarkable improvement in object classification and hence won the ImageNet ILSVRC challenge in 2012 [105]. ImageNet dataset is considered to be the most challenging collection of images in the computer vision community. The Deep CNN of [105] has later become famous as the AlexNet.

Since 2012, different variants of Deep CNNs have pushed the state-of-the-art even further in image classification. ZF Net [108] is an improvement over AlexNet with tweaked architecture hyper-parameters which has won the ImageNet ILSVRC challenge in 2013. GoogleLetNet [109] has developed an *Inception* module that dramatically reduced the number of parameters compared to AlexNet and they have introduced average pooling instead of the fully connected layers at the top of CNN that eliminated a large number of parameters without affecting performance. It has won the ImageNet ILSVRC challenge in 2014. VGGNet [110] showed that the depth of Deep CNNs is the key to their success, and it is the runner up in the ImageNet ILSVRC challenge in 2014. ResNet [111] is the winner of ImageNet ILSVRC challenge in 2015. It employed skip connections and batch normalization that acted as a regularizer and achieved training in fewer steps [112].

Apart from image classification deep networks have found applications in problems that have a temporal dependency in terms of Recurrent neural Networks (RNNs)

[113]. RNNs have applications in language modeling [114], speech recognition [115] and machine translation [116], but these are beyond the scope of the proposed work.

Supervised deep learning using convolution neural networks (CNN) (e.g., [117]) is proposed in the past. Tasks that fall within the paradigm of supervised learning are pattern recognition (also known as classification) and regression (also known as function approximation). The supervised learning paradigm is also applicable to sequential data (speech and gesture recognition). This can be thought of as learning with a "teacher", in the form of a function that provides continuous feedback on the quality of solutions obtained thus far.

Deep Belief Network (DBN) [118] and Autoencoder [119] are described in the literature as unsupervised deep learning. The core difference from supervised learning is the unavailability of the information about the desired output. Recently, it is shown in [120] that using unsupervised pre-training followed by limited supervised fine-tuning can build high level, class specific feature detector from unlabeled data. Competitive results can be achieved without doing labor intensive labeling of dataset for supervised algorithms. The researchers have achieved 70 % relative improvement over highest other result on ImageNet dataset [105].

2.3.2. Deep Convolution Neural Network. Before reviewing a deep convolutional neural network (Deep CNN), it is necessary to know what a simple neural network is. An artificial neural network (NN) is a feature extractor and a classifier. Artificial NN gets its motivation from human brain which consists of billions of neurons that are interconnected. Each brain neuron computes an accumulation of its inputs and decides its output state.

The structure of a NN consists of several components that are interconnected and organized in layers. These components are called artificial neurons (ANs). In general, a non-linear AN computes a weighted sum of its inputs and then outputs a non-linear function of the weighted sum. These weights can be considered as feature detectors. Each artificial neuron may be regarded as a simple classifier, which has limitations for complex non-linear problems. However, an interconnection and stacking of such simple classifiers yields to highly non-linear decision boundaries that can address

complex classification tasks. Hence, a number of ANs are interconnected to form a NN in order to overcome the limitations of simple isolated classifiers. AN can be a perceptron or a logistic regression unit, which are themselves standalone classification algorithms.

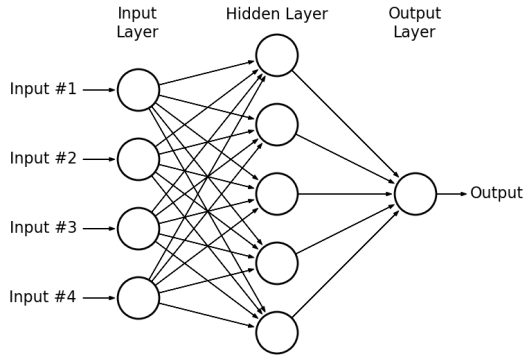

Figure 6: A typical neural network; here the input feature vector is 4 dimensional. The inputs are connected to a hidden layer of artificial neurons (ANs). The last layer consists of one AN, whose inputs are the weighted outputs of the ANs of the hidden layer (Image courtesy of citeNNim1).

A typical neural network is shown in Figure 6. It consists of three layers: an input layer, a hidden layer and the output layer. Hence, this is a 3-layered network. The hidden and output layers consist of artificial neurons. A neural network has two modes of operation, a *feed-forward* mode where the information progresses from the input layer to the output layer and *backpropagation* where the output of the network is matched to the true outputs and the error is propagated back into the network to adjust the weight, so that the error at the output layer is minimized [121].

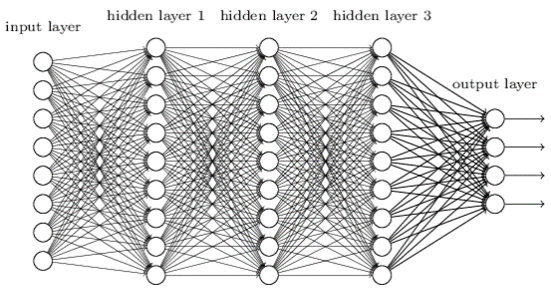

Figure 7: A deep neural network (image courtesy of [122])

A Deep Neural Network (DNN) is in principal a multilayered neural network. An example of a deep neural network is shown in Figure 7. However, the problem in naively stacking layers of hidden layers of neurons is that the number of weights/-parameters to learn in the network increases exponentially which may require a huge amount of labeled data. However, structure in the images can be exploited by using the information of correlation of the adjacent pixels in the image. Hence, fully connected layers can be replaced with layers of neurons that only connect adjacent pixels (or outputs from adjacent neurons). If their weights are shared across all image locations, it is essentially convolving the image with these weights.

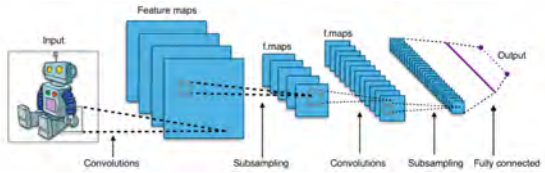

Figure 8: Structure of a CNN model

Similarly convolution can be applied on the image with different sets of weights. The results are then sub-sampled and maxpooled/average-pooled. Then, at the end, one or two fully connected hidden layers are used to increase the expressive power of the network. Such a paradigm makes the network invariant to translation in images and, with multiple layers, may learn complex representations. Such networks are termed as Convolutional Neural Networks (CNNs) [123]. A typical CNN is shown in Figure 8.

Deep CNNs, similar to DNNs, have several layers of convolution, pooling and local contrast normalization (shown to introduce brightness invariance [105]) stacked on top of each other and have one or two fully connected hidden layers [[105], [108], [109]]. The deep CNN used in [105] is shown in Figure 9. The other architectures are variants of this architecture. These can be trained by the backpropagation algorithm with stochastic gradient decent.

NN is generally restricted to few layers; for example, three layers; whereas, DNN has considerably more layers, which describes the term deep [124]. Each successive layer in DNN uses the output from the previous layer as input. The algorithms may be supervised or unsupervised and applications include pattern analysis (unsu-

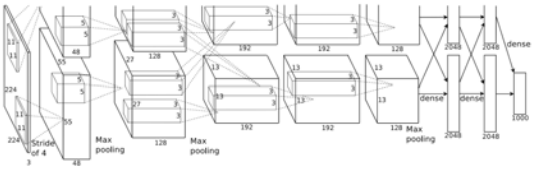

Figure 9: Deep CNN model used in [105]

pervised) and classification (supervised). The deep architecture allows the system to learn to represent features by themselves based on the nature of the data, rather than the subjective nature of human perception. This DNN architecture proved to achieve state-of-the-art in various computer vision tasks with little effort in tuning the model, including text recognition [125], object detection [126], object recognition [127], face recognition [128] and scene parsing/labelling [129].

Deep CNN is a supervised deep neural network which consists of a number of convolutional and subsampling layers optionally followed by different optimization and fully connected layers. The convolutional layer is the core building block of a CNN. Its filters have small receptive field but extends through the full input map. Each entry in the convolution layers output map is interpreted as an output of neuron that looks into the small region of the input map. Pooling layer is a non-linear, down sampling layer with options of several non-linear functions to implement. Pooling layer reduces the map size as well as adds translational invariance property to the CNN. Local contrast normalization layer operates at the output of the pooling layer. Its goal is to subtract mean and divide the standard deviation of the incoming neurons. This operation allows brightness invariance, which is useful for image recognition.

However, to avoid the deep CNNs from getting stuck in local minima, different strategies can be adopted during training; for example, unsupervised pre-training, drop-out, or increasing the training set by introducing noise and geometric transformation. In unsupervised pre-training, different network layers are initialized separately using autoencoders [120]. Autoencoders are neural networks that have the same inputs and output; however, these can learn interesting structures by introducing bottlenecks, such as making the number of hidden units lesser than the inputs or introducing sparsity constraints [119], [130]. This is done layer-by-layer using unlabeled images [120]. An-

other technique is dropout [105]; where the output of neurons is set to zero at random, which reduces *co-adaptation* of neurons.

2.3.3. Triplet based deep CNN model. Triplet-loss based DCNN model is used as described in [131], [108] to get low dimensional embeddings which can better represent subtle features like expressions due to its non-linear metric learning paradigm. In this method, the training dataset is arranged in the form of triplets (anchor, positive and negative). Anchor and positive images are from the same expression while the negative image has a different expression. Selection of positive and negative image pairs is based on α, which is a margin that is enforced between positive and negative pairs. The model learns 128 dimensional embeddings from the dataset and uses triplet loss function as shown in Equation 5 to bring matching face expression pair (anchor, positive) closer to each other and non-matching face expression pair (anchor, negative) farther from each other without any need of expression labels. Triplet loss for each triplet is calculated using Equation 3. This allows for clustering, not only unknown/-known classes, but also partition different intensities within same expression classes. An example of such triplet is shown in Figure 10.

$$|| f(x_i^a) - f(x_i^p) ||_2^2 + \alpha < || f(x_i^a) - f(x_i^n) ||_2^2 \tag{3}$$

$$\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \top \tag{4}$$

$$\sum_i^N \left[|| f(x_i^a) - f(x_i^p) ||_2^2 - || f(x_i^a) - f(x_i^n) ||_2^2 + \alpha \right] \tag{5}$$

Figure 10: Example of a triplet; anchor, positive and negative thumbnails (from left to right)

Triplet selection for training is a crucial task as it defines the overall performance of a model in an unconstrained environment. As the objective of the current research is to cluster unknown expressions as well as to separate different intensities within each expression, triplets are selected in such a way that they can separate weak expression classes well apart to improve model's effectiveness under unconstrained environment. There are different approaches which are used for triplet selection. In order to have the fast convergence of the triplet-loss, it is important to violate the triplet constraint in Equation 3. In other words, given x_i^a, it is required to select x_i^p (hard positive) as $argmax_{x_i^p} || f(x_i^a) - f(x_i^p) ||_2^2$ and x_i^n (hard negative), as $argmin_{x_i^n} || f(x_i^a) - f(x_i^n) ||_2^2$. However, it is not practically feasible to compute these constraints on complete dataset. Moreover, it may lead to poor training as poorly imaged faces will dominate the hard positives and hard negatives. In order to select triplets from labeled datasets, proposed technique in [131] is followed to select triplets from mini-batches while training. For the unconstrained dataset, Matlab based GUI is developed to manually select smattering number of triplets and further enhance to large number of triplets by taking the advantage of same expression of anchor and positive. If k is the number of triplets that exist for anchor i, then using loops and varying the corresponding positive and negative lists of length k triplets can be enhanced. The methodology is shown in Figure 11. Large number of triplets generated from this method fulfills the constraint in Equation 3.

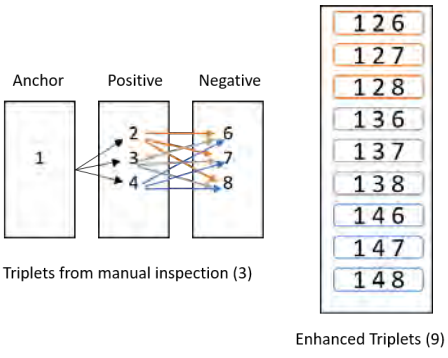

Figure 11: Proposed triplets enhancement technique is illustrated by an example of three manual triplets which are enhanced to nine triplets using this technique

Table 1 shows the architecture of the deep CNN model training expression models. Stochastic Gradient Descent (SGD) with standard backprop [132] is implemented

Table 1: Architecture of NN2 inception model [131] used for learning embeddings for facial expressions. The pooling is always 3×3 (aside from the final average pooling) and in parallel to the convolutional modules inside each Inception module.

layer	**size-in**	**size-out**	**param**	**FLOPS**
conv1	$224 \times 224 \times 3$	$112 \times 112 \times 64$	9K	119M
pool + norm	$112 \times 112 \times 64$	$56 \times 56 \times 64$	0	
rnorm1	$56 \times 56 \times 64$	$56 \times 56 \times 192$	115K	360M
inception(2)	$56 \times 56 \times 192$	$28 \times 28 \times 192$		
norm + pool	$28 \times 28 \times 192$	$28 \times 28 \times 256$	164K	128M
inception(3a)	$28 \times 28 \times 256$	$28 \times 28 \times 320$	228K	179M
inception(3b)	$28 \times 28 \times 320$	$14 \times 14 \times 640$	398K	108M
inception(3c)	$14 \times 14 \times 640$	$14 \times 14 \times 640$	545K	107M
inception(4a)	$14 \times 14 \times 640$	$14 \times 14 \times 640$	595K	117M
inception(4b)	$14 \times 14 \times 640$	$14 \times 14 \times 640$	654K	128M
inception(4c)	$14 \times 14 \times 640$	$14 \times 14 \times 640$	722K	142M
inception(4d)	$14 \times 14 \times 640$	$7 \times 7 \times 1024$	717K	56M
inception(4e)	$7 \times 7 \times 1024$	$7 \times 7 \times 1024$	1.6M	78M
inception(5a)	$7 \times 7 \times 1024$	$7 \times 7 \times 1024$	1.6M	78M
inception(5b)	$7 \times 7 \times 1024$	$7 \times 7 \times 1024$		
avg pool	$7 \times 7 \times 1024$	$1 \times 1 \times 1024$	131K	0.1M
L2 norm	$1 \times 1 \times 1024$	$1 \times 1 \times 128$	0	
Total			7.5 M	1.6 B

in the model. The architecture is based on the inception model of Szegedy et al. [109]. These networks use mixed layers that run various convolutional and pooling layers in parallel and concatenate their responses. It is found that the number of parameters can be reduced to 20 times and also has the potential to reduce the number of FLOPS required for comparable performance as compared to DCNN proposed by [108]. Deep CNN model has an overall of 17 layers as shown in Table 1. It has a total of 7.5M million parameters and requires around 1.6 billion floating point operations per second (FLOPS) per image. The novelty of the network lies in the use of inception module which dramatically reduces the number of parameters in the network. It also uses Average Pooling instead of Fully Connected layers that reduce the number of parameters as well as improve robustness to spatial translation. Since convolutional filters can learn linear functions of their inputs, there is a need to have more complex filters which have more learning capabilities and abstraction power. Szegedy et al. used multi-layer perceptrons to connect convolutional layers which are mathematically equivalent to 1×1 convolutions and thus fit within the CNN architecture. Convolutional layers reduce

the dimensionality of the feature space while keeping information intact. Convolutions with different filter sizes are applied in parallel to recover both local feature via smaller convolution and high abstracted features with larger convolutions. One of the Inception modules used in Table 1 is shown in Figure 12.

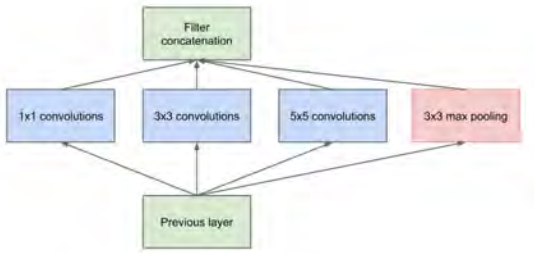

Figure 12: Inception module used in the CNN architecture (image courtesy of [133])

2.4. Clustering

Clustering is the process of grouping a set of data points into classes of similar objects. A cluster is a group of data points that are similar to one another within the same cluster and dissimilar to the objects in other clusters. For the problem statement of the current research work, K-Means and consensus clustering are used to cluster the extracted features into different classes.

2.4.1. K-means. K-means is the most popular hard deterministic clustering algorithm. It is also known as "Lloyd's algorithm" in computer science community. It is a method for vector quantization. It is commonly used in cluster analysis in problems related to machine learning, data mining and others. The objective of this algorithm is to cluster n observations with d dimensions into k clusters where ($k \leq n$) is based on a distance metric selected for the problem as shown in Figure 13. Cluster assignment to any observation is based on the nearest mean. Euclidean distance is commonly used as a distance metric, so the equation is as follows:

$$||x^t - m_i|| = min_j ||x^t - m_j|| \tag{6}$$

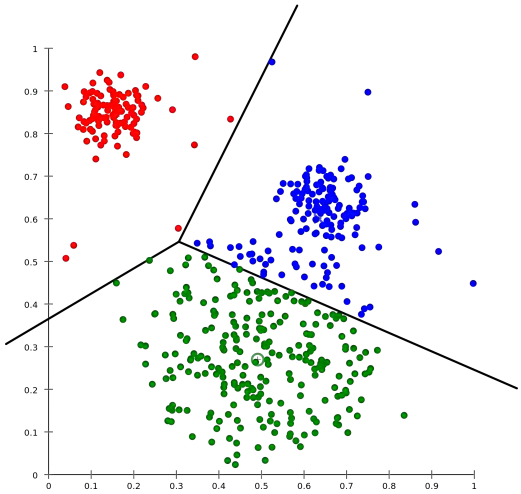

Figure 13: An example with k=3 means

given reference vector m_j, j=1 ...k. Instead of the original data, m_j reference vectors are used as a codebook vectors or code words, as this is an encoding /decoding process as shown in Figure 14.

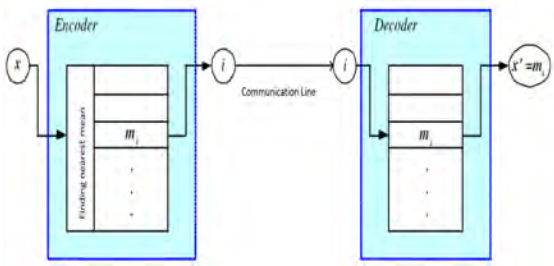

Figure 14: Given input x, the encodes sends the index of the closest code word and decoder generated the code word with the received index $x^{'}$ with error $||x^{'} - x||^2$, adopted from [134]

Mahalanobis distance metric is also used in place of Euclidean. If the covariance of the natural groupings are not identity matrices, they instead have some elliptical representation. However, some drawbacks of K-means algorithm are as follows:

- Clustering of non-clustered dataset as division is primarily dependent on given k.
- It is sensitive to scaling and it is set to the assumption that all variables have equal variances.

- Sometimes it may get stuck to local minimum even on perfect datasets with clear separation.

  The common technique used to calculate K-means is an iterative refinement approach. Total reconstruction loss calculation in K-means algorithm is shown in Algorithm 1.

Algorithm 1 K-means clustering algorithm

1: Repeat For all $x^t \in X$
2: $b_i^t \quad \{\ 1, \quad if\ \|x^t - m_i\| = min_j \left\|x^t - m_j\right\| \quad 0, \quad Otherwise\ \}$

3: For all m_i, i=1 ...k

4: **while** m_i is converging **do**
5: $\quad m_i \leftarrow \sum_{t=1}^{X} b_i^t x^t / \sum_{t=1}^{X} b_i^t$

6: **end while**
7: $E(\{m_i\}_{i=1}^{k} / X) = \sum_t \sum_i b_i^t \|x^t - m_i\|$

2.5. Consensus Clustering

In statistical data analysis, clustering is a common technique for dimensionality reduction and grouping data in different clusters. It is also used in a number of computer vision fields, such as data mining, machine learning and pattern classification. Consensus clustering is simply the clustering technique in which objects from the same cluster have more similarities than the objects from different clusters. Similarity can be treated as a distance measure within different objects. It is also known as aggregation of partitions, refers to a scenario in which input is the number of clusters (partitions) which are obtained from a dataset. Furthermore, it is required to get a consensus clustering which is thought of a better representation of clustering than the pre-defined group of partitions in terms of scalability, stability, parallelization and robustness. Consensus clustering can be thought of as an adaptive clustering data from the same dataset which are coming from different algorithms or multiple runs of the same algorithm, which are used to form co-association matrix. Cluster position of each image is compared to rela-

tive partition/cluster position of all other images from the dataset. Output is 1 if both lie in the same cluster and is 0 if otherwise. Then, sum of all combinations of every single image is normalized to (0-1) intensity level by dividing it by the sum of total number of images. Co-association matrix is *nxn* dimensions and represents similarity between each image. Spectral clustering [135] is a method of finding clusters using top eigenvectors of a matrix derived from the distance between points. Consensus clustering for unsupervised learning is shown in Figure 15.

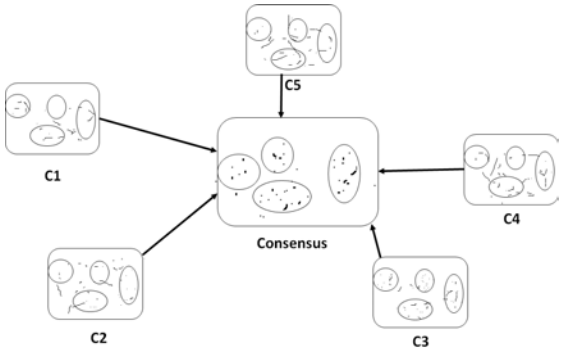

Figure 15: Consensus clustering based on different partitions from the same dataset is shown to illustrate the process

Traditional clustering algorithms use limited visual descriptors as basic features to cluster data into different groups. Some algorithms use intensity or pixel values of the image data as feature set which increases the complexity of clustering. There is no obvious similarity measure (distance measure) for clustering, and it needs to be defined. Nonetheless, this is not simple, especially in multidimensional space. Due to these potential shortcomings, interpretation of results becomes somewhat problematic chiefly when there is no information about the number of clusters [136]. Consensus clustering is a two-step method; multiple folds of cluster formation on a dataset (can be raw pixels or some other features such as LBP and PCA) and then consensus clustering on those clusters to get an aggregated set of clusters which have better representation. Consensus clustering is usually performed on clusters/partitions formed by any clustering method such as K-means, DBSCAN or Agglomerative algorithms as they cluster the spatial information in an efficient manner.

In order to classify facial expressions in the wild with no prior information of the number of expressions, consensus clustering algorithm is selected either with raw pixels, pure embeddings, LDA transformed or PCA transformed features from either raw pixels or DCNN based non-linear embeddings as these have the best purity levels compared to other proposed algorithms.

Chapter 3: Databases

Three datasets are used; one unconstrained and two constrained datasets. Datasets taken under lab-controlled environment are Multi-PIE and MMI while unconstrained dataset is made by open license videos from YouTube. Details of each dataset are given below.

3.1. Multi-Pie Dataset

The CMU Multi-PIE face database [137] consists of more than 750,000 images from 337 individuals which are recorded in up to four sessions over the span of five months. Subjects are imaged under 15 different poses and 19 illumination conditions but with limited facial expressions. This dataset contains images of happy, normal, surprised, disgusted and open mouth expressions. In addition, high resolution frontal images are acquired. In total, the database contains more than 305 GB of face dataset.

Multi-PIE has images of positive, negative, surprised and normal expression classes taken under constrained environment. It can be easily distributed into different expressions based on ground truth provided with each image. Matlab code is used in which one of the two expression image lists is passed to anchor and positive vectors and the other expression image list to negative vector. Then, triplets are generated using nested loops in which outermost loop is on anchor vector and innermost is on negative list. This process is repeated for all partitions which are described in Table 2. Overall, 2,00,000 triplets are generated for each 2-expression partition which are divided into train and validation sets in 2:1 proportion. An example of an image triplet from Multi-PIE dataset is shown in Figure 16.

Figure 16: Example of a triplet from PIE dataset

3.2. MMI Database

MMI facial expression database is developed by Maja Pantic, Michel Valstar and Ioannis Patras in 2002 as a resource for building and evaluating facial expression recognition algorithms [138]. Database consists of over 2900 videos and high-resolution images of 75 subjects. All videos are fully annotated for the presence of AUs and partially coded on frame level whether AU in the specified frame is neutral, onset, apex or offset phase of the expression. In addition to six basic emotions, the MMI database contains prototypical expressions and expressions with a single FACS Action Unit (AU) activated, for all existing AUs and many other Action Descriptors. Recently recordings of naturalistic expressions have been added too. Example images from the dataset are shown in Figure 17.

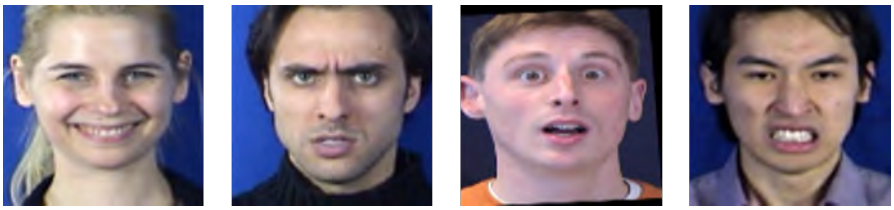

Figure 17: Sample images from MMI database

3.3. Face Dataset from YouTube Videos

YouTube dataset is created with the motivation to test unsupervised facial expression clustering in the wild. For this, videos from YouTube are used under creative common license. In order to acquire a large dataset with several expressions, more than 150 videos are downloaded from different genres with the aim to get multiple faces with multiple expressions under unconstrained environment. Python library is used to automatically download lists of specified videos from YouTube. The next step is face detection and extraction. Dlib library's [139] face detector is used to detect and extract bounding boxes containing faces from all frames of the videos. Faces are aligned using facial key-points, mean and variance normalized and stored in 100×100 dimensions. Experimental setup and outcomes of the proposed algorithms are explained

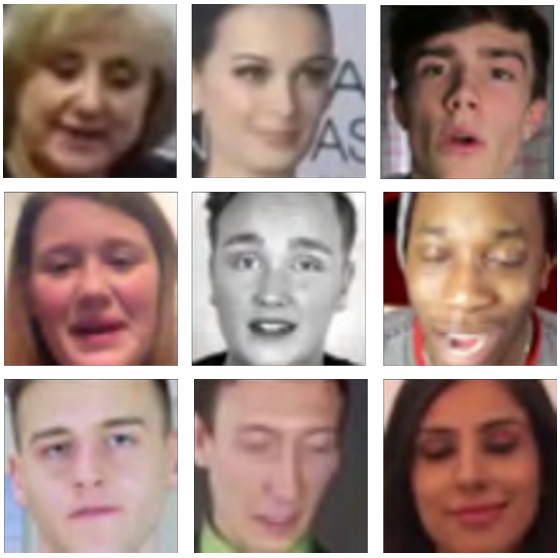

Figure 18: Images from the dataset created from YouTube videos

in experimental setup and results section. Some faces from the dataset are shown in Figure 18.

YouTube image dataset are created from creative content videos (open license videos) without any information of expressions or any ground truths. For that, Matlab Graphical User Interface (GUI) is proposed as shown in Figure 19. Images are

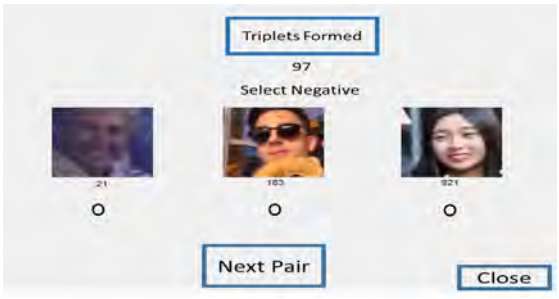

Figure 19: Matab GUI used to form expression based triplets for the youTube dataset

randomly selected to remove any inclination towards individuality features. To aid randomization and to increase representation of each image in triplets, the image list is divided into five successive overlapping sub-lists. Afterwards, combinations for each sub-list are formed using three samples per observation and selected 5000 combinations from each sub-list. Images are loaded and displayed on the GUI as shown in Figure 19. Triplets for both extreme cases are skipped to identify if all three images had same ex-

pression or all three had different expressions. The negative thumbnail is selected from the image triplet using the information that anchor and positive must have the same expression. Presently displayed triplet are added to the triplet list after selecting negative image from the triplet, and a new image triplet is loaded when the user clicked on 'Next pair' button. The same process is repeated over all combinations. A number of 19,705 triplets are collected after manually inspecting each triplet from the pool of 25,000 triplets. Manually selected triplets are handful compared to DCNN dataset requirement for better training; therefore, triplets are enhanced using the proposed triplet enhancement method as shown in Figure 11. Triplets are enhanced by using the knowledge of having positive and negative thumbnails of the triplet from the same expression class.

Chapter 4: Experiments and Results

A similar pipeline for all the three datasets is followed. Face detection and alignment is done using key-point detection [139] and then resizing them to 100×100 pixels. Dimensionality reduction is done by learning linear or non-linear embeddings and clustering the expression features into different clusters. An expression class is excluded while learning the embeddings and is then included in the test partition during clustering facial expressions.

The motivation behind reducing dimensions is two-fold. Firstly, clustering is easier in reduced dimensions. Secondly, subspace is intended which has faces from same expressions closer to each other and those with different expressions, farther away. Three methods are used to achieve the set goals: Principal Component Analysis (PCA) [47], Linear Discriminant Analysis (LDA) [81] and learn embeddings via a Deep CNN that minimizes the triplet loss [97].

Positive, negative and surprised facial expression images are used. Matlab code is used to automatically generate a total of 2,00,000 triplets for each 2-expression partitions. These triplets are further divided into train and validation sets in 2:1 proportion.

MMI dataset is also a constrained database. However, in addition to six basic emotions, it has some prototypical and only one AU activated expressions. Each video is labeled; therefore, triplets are generated in a similar fashion as for Multi-PIE dataset. YouTube dataset is highly unconstrained in terms of subjects, expression intensities, pose, lighting, etc. as videos are taken from different genres. In order to remove biasness in results due to any long duration video, 30 face images are taken from each video, and a pool of 3000 images is constructed. Matlab GUI is used to manually select a handful number of triplets which are automatically enhanced to overall 1,70,000 using the technique demonstrated in Figure 11.

Each dataset is divided into two independent partitions having 60-40% representation of the dataset. Therefore, two sets of experiments are performed; training on 60% partition and test on 40% partition and vice versa.

In order to select the optimal dimensions for PCA, experiments are performed on raw features of Multi-PIE dataset with different PCA dimensions in the range of

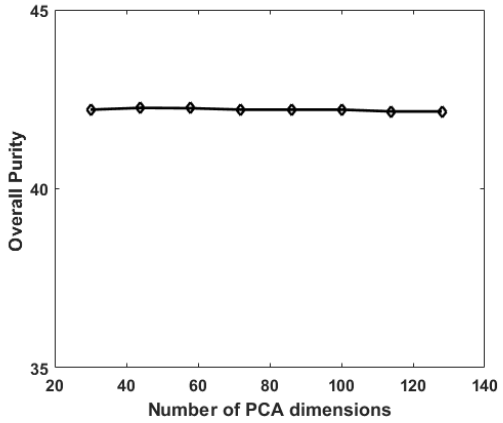

Figure 20: Purity levels are plotted against different PCA dimensions for the experiment are performed on raw features of PIE dataset with PCA-Kmeans.

30-128. PCA dimensions are plotted versus purity levels as shown in Figure 20, and then 100 dimensions are selected for PCA dimensionality reduction algorithm for the experimental setup as they ensured high purity level with **90+ %** proportion of variance. Supervised LDA is also used to transform the dataset into (c-1) dimensions, where c is the total number of classes in the training dataset. Triplet-based DCNN is used to extract non-linear transformation of the dataset. Learning rate of the model is set to 0.01. All layers are initialized with random values. Decrease in the total loss slowed down drastically after 100th iteration of the training which encouraged the researcher to train the model with a small number of iterations. A number of 120 iterations per each experiment is used. Each iteration is further sub-divided into 1500 batches with a mini-batch size of 30 triplets. The margin α is set to 0.2.

K-means and consensus clustering algorithms are used to cluster the extracted features into n clusters defined by purity levels or NMI values for Multi-PIE and MMI datasets. NMI values using Equation 7 are maximum when $n = 3$, and $n = 12$ are used for Multi-PIE and MMI datasets, respectively, which is justified as n is equal to the total number of expressions in the test datasets. For YouTube dataset, few possibilities for the number of clusters are practiced on a validation subset, and a number which gave the best subjective quality of clustering is chosen.

NMI is a measure of mutual information about two variables. It quantifies the amount of information obtained about the random variable given the other random vari-

able as shown in Equation 7. NMI is always a number between 0 and 1.

$$NMI(\Omega,\Theta) = \frac{I(\Omega;\Theta)}{|[H(\Omega)+H(\Theta)]/2|} \tag{7}$$

where $I(\Omega;\Theta) = \sum_k \sum_j \frac{|w_k \cap c_j|}{N} log \frac{N|w_k \cap c_j|}{|w_k||c_j|}$ and $H(\Omega) = -\sum_k \frac{|w_k|}{N} log \frac{|w_k|}{N}$

The prime goal is to cluster unknown expressions in the wild. For that, the algorithm is first tested on constrained datasets with known expression classes. While training, an expression class is left in each experiment and included while testing the model to quantify the strength of the proposed semi-supervised method in terms of purity of clustering. To compare the performance of constrained and unconstrained datasets, all datasets are divided into non-overlapping train and test datasets to remove any biasness towards results. Faces are extracted from videos/images. Then, they are aligned and mean and variance are normalized to reduce illumination variations. Raw pixels and embeddings from deep CNN are used as input features while unsupervised PCA and supervised LDA are used for dimensionality reduction. Finally, clustering of facial expressions is performed using K-means and consensus clustering.

4.1. Multi-PIE dataset

Instead of using universal naming, positive (happy), negative (sad, disgust) and surprise facial expressions are used. For the experiment, 9853 front pose images are used from a total of 335 individuals. Three randomizations, with two partitions each are generated. Subjects in both partitions from any randomization are independent. The reason behind partitioning is to evaluate the algorithm on different sets of individuals to avoid inclination towards any expression in the dataset. Distribution of different partitions used in the experiments are shown in Table 2. The idea behind making multiple subject independent folds is to get rid of any bias in the results.

An expression from training dataset is excluded in each DCNN model training and tested on all three expressions from the test partition. This process is repeated for all possible combinations.

Table 2: Distribution of three randomizations (Random 1, Random 2, Random 3) of Multi-PIE dataset into two partitions to ensure the strength of the algorithm. Both partitions in one randomization are subject independent

	Random 1		**Random 2**		**Random 3**	
Expression	**P1**	**P2**	**P1**	**P2**	**P1**	**P2**
Positive	1699	2537	1758	2478	1822	2414
Surprise	707	1071	767	1011	760	1018
Negative	1728	2111	1591	2248	1664	2175
Total	4134	5719	4116	5737	4246	5607

Experimental setup for each partition in each fold of the dataset is done as follows:

- Training without positive expression and test on all three expressions from the other partition of the same fold
- Training without negative expression and test on all three expressions from the other partition of the same fold
- Training without surprise expression and test on all three expressions from the other partition of the same fold

The following sets of algorithms are used in the test on the dataset:

- Embeddings + K-means
- Embeddings + PCA + K-means
- Embeddings +LDA +K -means
- Embeddings + consensus clustering
- Embeddings + PCA + consensus clustering
- Embeddings + LDA + consensus clustering
- Raw Pixels + K-means
- Raw Pixels + PCA
- Raw Pixels + LDA
- Raw Pixels + consensus clustering
- Raw Pixels + PCA + consensus clustering
- Raw Pixels + LDA + consensus clustering

Two models are used for the aforementioned tests with input features (raw pixels or embeddings from trained DCNN model). These models are: PCA with K-

Table 3: Weighted average purities of all experiments on three randomized partition 1 of Multi-PIE dataset. Features used in the experiment are raw pixels (R) and embeddings (E) from triplet based CNN model. PCA (P) and LDA (L) are used for dimensionality reduction. Clustering algorithms are K-means (Km) and consensus clustering (Cons).

Features / Classifiers	**Trained with no positive**	**Trained with no negative**	**Trained with no surprise**
R-P-km	42.48	42.48	42.48
R-P-Cons	42.49	42.32	42.59
R-Cons	42.47	42.33	42.47
R-L-km	54.01	56.62	63.23
R-L-Cons	55.68	55.81	65.05
E-Km	65.59	61.94	65.99
E-L-km	67.73	59.97	67.11
E-P-km	65.75	61.98	65.98
E-Cons	66.35	**63.29**	**63.79**
E-L-Cons	**67.75**	59.94	67.36
E-P-Cons	66.24	63.29	63.91

Table 4: Weighted average purities of all experiments on three randomized partition 2 of Multi-PIE dataset. Features used in the experiment are raw pixels (R) and embeddings (E) from triplet based CNN model. PCA (P) and LDA (L) are used for dimensionality reduction. Clustering algorithms are K-means (Km) and consensus clustering (Cons).

Features / Classifiers	**Trained with no positive**	**Trained with no negative**	**Trained with no surprise**
R-P-km	43.54	43.54	43.54
R-P-Cons	43.48	43.46	43.47
R-Cons	43.47	43.47	43.47
R-L-km	44.41	45.79	59.03
R-L-Cons	44.20	45.00	61.53
E-Km	63.11	57.69	66.39
E-L-km	66.37	59.85	68.03
E-P-km	63.21	57.65	66.36
E-Cons	64.06	58.57	64.98
E-L-Cons	**66.38**	59.82	**68.11**
E-P-Cons	64.09	**59.42**	64.88

means/consensus clustering and LDA with k-means/consensus clustering, and the one with overall better accuracy is picked based on purity.

Table 5: Triplet loss model training on negative and surprise expressions from partition 2 of the Multi-PIE dataset and test performed on embeddings of all expression classes from partition 1 extracted using aforementioned DCNN model. Clusters are showed from PCA (P), LDA (L) with k-means (Km) on raw pixels and PCA (P), LDA(L) with K-means (Km) and consensus clustering (Cons) on embeddings.

R-L-Km	**Cluster1**	**Cluster2**	**Cluster3**
Positive	1145	527	865
Surprise	425	556	90
Negative	1019	475	617
Purity	44.23	35.69	**55.03**
R-P-Km	**Cluster1**	**Cluster2**	**Cluster3**
Positive	904	772	861
Surprise	419	327	325
Negative	760	650	701
Purity	43.4	44.14	**45.63**
E-Km	**Cluster1**	**Cluster2**	**Cluster3**
Positive	350	456	893
Surprise	0	639	68
Negative	1194	42	492
Purity	77.33	56.2	**61.46**
E-L-Km	**Cluster1**	**Cluster2**	**Cluster3**
Positive	888	477	334
Surprise	86	0	621
Negative	429	1290	9
Purity	**63.29**	73.01	64.42
E-P-Km	**Cluster1**	**Cluster2**	**Cluster3**
Positive	891	352	456
Surprise	69	0	638
Negative	489	1196	43
Purity	**61.49**	77.26	56.11
E-L-Cons	**Cluster1**	**Cluster2**	**Cluster3**
Positive	334	477	888
Surprise	621	0	86
Negative	9	1290	429
Purity	64.42	73.01	**63.29**
E-P-Cons	**Cluster1**	**Cluster2**	**Cluster3**
Positive	1085	265	349
Surprise	143	564	0
Negative	529	6	1193
Purity	**61.75**	67.54	77.37

Purity is a simple and transparent evaluation measure. To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned instances and divide by the total number of instances N as shown in Equation 8.

$$Purity(\Omega, \Theta) = \frac{1}{N} \sum max_j |w_k \cap c_j| \tag{8}$$

where $\Omega = w_1, w_2, ..., w_k$ is the set of clusters and $\Theta = c_1, c_2, ..., c_j$ is the set of classes. In this case, w_k represents the expression clusters formed by k-means while c_j corresponds to expression classes (positive, negative and surprise).

Initiation of K-means algorithm is random; therefore, to avoid any bias in the results, K-means clustering is performed on five unique random states. Semi-supervised clustering on raw images is the initial approach towards unsupervised clustering of unknown expressions. LDA with raw pixels gives better purity values compared to clustering on raw pixels. It is due to the subtleness of expression features that unsupervised K-means and PCA clustered subject dependent features rather than expressions as shown in Tables 3 and 4.

Each value in Tables 3 and 4 is the weighted average values using partitions from three different randomizations of the dataset. Weighted average purities on embeddings from DCNN with consensus clustering gives the best results with much better class separation for the unknown expression compared to corresponding tests on raw features as shown in Tables 3 and 4. Consensus clustering on embeddings with PCA dimensionality reduction gives **23 %** higher overall clustering purities on average compared to the same clustering and dimensionality reduction techniques when applied on raw features. Table 5 shows the clustering distribution for different experiments performed on raw features and non-linear embeddings from DCNN in which positive expression is left during model training and included it while clustering on test dataset. The trend remained same when purity per cluster is considered instead of overall test purity. It is concluded from the comparative study on Multi-PIE dataset that embeddings from DCNN model (trained with an expression left) give much better unknown expression class separation compared to linear embeddings on raw pixel dataset.

4.2. MMI Dataset

MMI dataset contains six basic expressions (happiness, surprise, disgust, anger, fear and sadness), neutral and five prototypical expressions. Five models are trained, each without one of the expression from the dataset. It is already established that embeddings with consensus clustering separate unknown expression class with far better purity compared to its corresponding test on raw features in Tables 3 and 4. Therefore, two tests are performed. These tests are embeddings-PCA-consensus clustering and raw pixels-PCA- (K-means). The number of clusters in each experiment is set to 12 based on purity level calculated using Equation 8. Results from five different experiments with an unknown expression class (surprise, fear, sadness, happiness and one prototypical expression) respectively are shown in Tables 7 and 8 using DCNN embeddings and raw features respectively as input features. The distribution of data is only showed in the unknown expression cluster in Tables 7 and 8. Unknown expressions are clustered with high purity level in each experiment using embeddings from DCNN; whereas, raw features based experiments clustered unknown expression with low purity level. The trend is consistent in the overall purity levels for each experiment as shown in Table 6.

it is worth noting that the purity for training without P-5, for the cluster that is assigned to P-5 is low, but only 141 images for P-5 and 121 out of those images get clustered in the same place. Better results could be obtained if more images for this expression are present.

Table 6: Results of all experiments performed on MMI dataset. Features used in the experiment are raw pixels (R) and embeddings (E) from triplet-based DCNN. PCA (P) is used for dimensionality reduction. Clustering algorithm are K-means (Km) and Consensus clustering (Cons). Excluded expressions in each model are (from left to right); surprise, fear, sadness, happiness and prototypical expression (P-5).

Features/ Classifiers	**Trained with no surprise**	**Trained with no fear**	**Trained with no sad**	**Trained with no happy**	**Trained with no P-5**
R-P-Km	27.62	27.82	27.37	27.41	28.52
E-Km	88.46	**87.84**	80.35	82.27	74.51
E-P-Km	88.53	86.26	80.41	81.54	74.13
E-P-Cons	**90.22**	86.12	**80.94**	**82.31**	**75.75**

Table 7: Results on Embeddings MMI dataset from DCNN with PCA-consensus clustering. Each column represents the cluster for the unknown expression with maximum purity level from each experiment. Excluded expressions in each model are (from left to right); surprise, fear, sadness, happiness and prototypical expression (P-5).

Expressions	**Trained with no surprise**	**Trained with no fear**	**Trained with no sad**	**Trained with no happy**	**Trained with no P-5**
Angry	1	2	5	1	1
Disgust	4	1	13	18	0
Fear	29	**490**	7	4	1
Happy	13	23	10	**609**	6
Sad	7	8	**757**	8	16
Surprise	**663**	36	7	0	9
P-1	0	0	0	0	2
Neutral	0	0	0	0	0
P-2	0	0	0	0	83
P-3	0	0	0	0	0
P-4	0	0	0	1	21
P-5	5	2	1	0	**120**
Purity	91.83	87.19	94.63	95.01	46.33

Table 8: Results on raw pixels MMI dataset with PCA-(K-means). Each column represents the cluster for the unknown expression with maximum purity level from each experiment. Excluded expressions in each model are (from left to right); surprise, fear, sadness, happiness and prototypical expression (P-5).

Expressions	**Trained with no surprise**	**Trained with no fear**	**Trained with no sad**	**Trained with no happy**	**Trained with no P-5**
Angry	86	156	58	0	42
Disgust	86	132	59	40	52
Fear	25	**197**	80	0	26
Happy	101	170	120	**223**	72
Sad	55	138	**157**	0	72
Surprise	**133**	192	73	0	47
P-1	0	16	37	0	0
Neutral	0	0	0	0	91
P-2	0	0	0	0	40
P-3	0	0	0	0	56
P-4	11	0	0	0	18
P-5	23	0	0	0	81
Purity	25.58	19.68	20.55	84.79	15.24

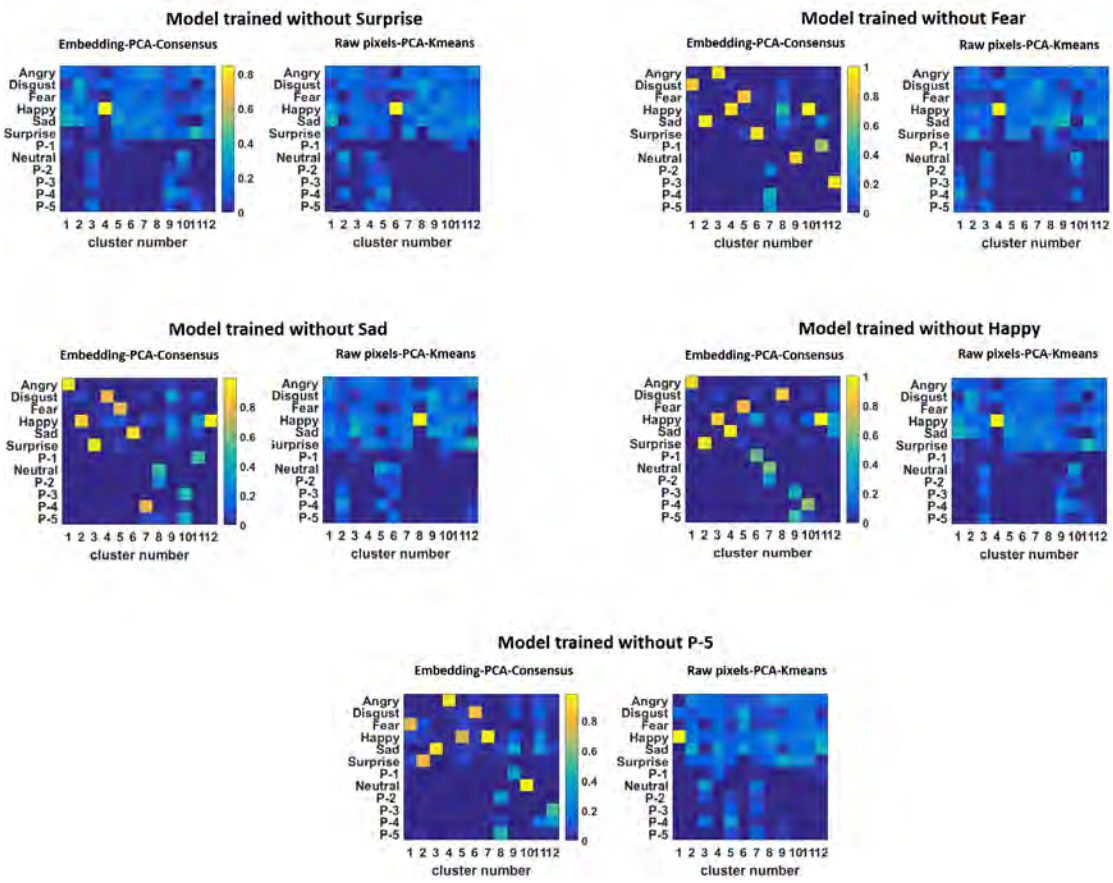

Figure 21: Comparative analysis between raw features and DCNN based embeddings on MMI datasets from experiments when the left expressions are surprise, fear, sadness, happiness and prototypical expression (P-5). Representation of each expression in clusters can be visualized by the color intensity chart associated with it.

Figure 22: Sample images are shown from sad and prototypical expression (P-5) clusters using model trained without P-5 expression. First two row images are from P-5 expression cluster and next two row images are from sad expression cluster.

Table 6 shows the results of all experiments performed on MMI dataset on five test cases, each with one expression class left during training (surprise, fear, sadness, happiness and prototypical expression P-5, respectively). Experimental setup is reduced to raw features-PCA-(k-means) and embeddings-(k-means), embeddings-PCA-(k-means), embeddings-PCA-consensus clustering as it is already proved that non-linear embeddings from DCNN clustered unknown expressions are of much higher purity level compared to raw features. Purities in Table 6 present a stork increment when non-linear embeddings are used instead of raw features. Figure 21 gives a comparative analysis of all five experiments performed on MMI datasets between DCNN embeddings and raw features. Results show that embeddings from DCNN could cluster known as well as unknown expressions with high purity levels. K-means clustering on PCA components using raw features could only cluster happy expression which is shown in Table 8.

Prototypical expression (P-5) used in the experimental setup is the high intensity version of the sad expression, and clustering results on DCNN embeddings using trained model without P-5 expression ensure that the proposed method does not only cluster discrete expressions, but also separates different intensity versions of the same expression as shown in Figure 22.

4.3. YouTube image dataset

Dataset is created using 150 YouTube videos and comprised more than 1.7 million images. There are numerous frames in each video where object's expression and pose remained unchanged; therefore, 3000 images are randomly selected from a set of image folders extracted from aforementioned video dataset. Our own triplet formation GUI, as discussed in section 3.3, is used. Generated set of triplets are used to train triplet-loss based CNN model. Remaining image folders are used to make a dataset of 2300 images with random subjects and expressions. Once the model is trained, embeddings on the test dataset are generated. Since the final number of expression classes are unknown, similarity matrix is computed using multiple cluster sizes in the range of 6-34. Consensus clustering is performed on the similarity matrix calculated

from various basic partitions formed using different parameters of the same algorithm. Consensus clustering is also performed on raw images to show the advantage of using semi-supervised method with triplet loss training.

Experiments are performed on raw images and embeddings from DCNN. The summary of the results is as follows:

- Images from all expressions are mixed up in all clusters as shown in Figure 23.
- Most individual images with different expressions are clustered in the same folder as shown in Figure 24.

Figure 23: Images from different expression classes lie in the same cluster when K-means clustering is performed on PCA components of raw features

- Consensus clustering on the embeddings from test dataset using YouTube model gives considerably better clustering of expressions. Examples from three clusters are shown in Figure 25. It is visually evident that different expressions are clustered in different partitions.
- Some individual images with completely different expressions are clustered into different expressions which shows the focus of trained model on expression features rather than physical appearance of the individual.
- There are some instances in which images from the same individual with minor difference in expressions are clustered together. It could be due to training

Figure 24: Images from an individual with different expressions in the same cluster when K-means clustering is performed on PCA components of raw features

on limited dataset and highly complex unconstrained feature space. This can be minimized in future work either by increasing training dataset or using more objective directed features.

Figure 25: Result of consensus clustering on test dataset embeddings using YouTube triplet loss model. First two rows are for cluster 1, next two rows are for cluster 2 and last two rows are for cluster 3. Note that, faces are taken from YouTube videos under complete unconstrained environment, therefore alignment has a strong impact on some images.

Tables 3 and 4 represent the weighted average purities of Multi-PIE dataset. Table 6 shows the comparative analysis of all experiments performed on MMI dataset based on the basis of purity values. It is evident from the purity values that the proposed method, consensus clustering with embeddings from triplet-loss based DCNN, can cluster unknown expression classes with significantly higher purity compared to other linear feature extraction methods. Based on the results, it is conjectured that this can be extended to cluster unknown expressions in the wild. Thus, some results on a completely unconstrained dataset that is downloaded from YouTube are also shown. Figure 25 shows the clustering of expressions into different clusters.

Chapter 5: Conclusion and Future work

A semi-supervised facial expression recognition algorithm is proposed with the motivation of clustering facial expression under unconstrained environment. Deep convolution neural network with triplet-loss training is applied on limited expressions as a metric learning paradigm to reduce complexity of the dimensions and to bring similar facial expressions closer to each other. CMU Multi-PIE, MMI Facial as well as on the own proposed YouTube dataset, which is entirely unconstrained in terms of expressions, subjects, pose, and illumination, are used to validate the proposed DCNN. The trained weights of the neural network model are used to compute embeddings on test dataset which are further refined and used with consensus clustering to cluster more expressions which are even non-existent in metric learning dataset. Results have shown that the proposed semi-supervised algorithm of triplet-loss based deep CNN model embeddings with consensus clustering can cluster unknown facial expressions in the wild with high purity level. Our approach of using DCNN based embeddings for test dataset to get better clustering is also proved by a significant purity level jump between embeddings based and raw features based results on the same set of algorithms. It is shown that the proposed work not only produced best clustering results on discrete expressions compared to other linear embeddings but also clustered expressions with different intensities. In order to cluster more unknown expressions with better purity level in the future, feature selection is to be modified. Since it is known that if the input to the Deep CNN has large number of features very large datasets are required to train them efficiently, there is a chance of over-fitting. Raw features are not the best input features because of the size of feature space and weak representation of required expression based features. Another future work can be training of the proposed work on large number of known expressions and test on further larger dataset in terms of expressions and then compare the trend of purity with more number of classes in training.

References

[1] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, 1999.

[2] S. Weerachai and M. Mizukawa, "Human behavior recognition via top-view vision for intelligent space," in *International Conference on Control Automation and Systems (ICCAS)*. IEEE, 2010, pp. 1687–1690.

[3] R.-H. Chen, R.-J. Lin, and P.-C. Yang, "The relationships between ecrm, innovation, and customer value-an empirical study," in *International Summer Conference of Asia Pacific on Business Innovation and Technology Management (APBITM), 2011 IEEE*. IEEE, 2011, pp. 299–302.

[4] M. Wimmer, B. A. MacDonald, D. Jayamuni, and A. Yadav, "Facial expression recognition for human-robot interaction–a prototype," in *International Workshop on Robot Vision*. Springer, 2008, pp. 139–152.

[5] J. F. Cohn, "Foundations of human computing: facial expression and emotion," in *Proceedings of the 8th International Conference on Multimodal Interfaces*. ACM, 2006, pp. 233–238.

[6] A. Pentland, "Socially aware, computation and communication," *Computer*, vol. 38, no. 3, pp. 33–40, 2005.

[7] M. Pantic, A. Pentland, A. Nijholt, and T. S. Huang, "Human computing and machine understanding of human behavior: a survey," in *Artifical Intelligence for Human Computing*. Springer, 2007, pp. 47–71.

[8] C. L. Lisetti and F. Nasoz, "Maui: a multimodal affective user interface," in *Proceedings of the Tenth ACM International Conference on Multimedia*. ACM, 2002, pp. 161–170.

[9] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724–736, 2007.

[10] A. Hanjalic and L.-Q. Xu, "User-oriented affective video content analysis," in *IEEE Workshop on Content-Based Access of Image and Video Libraries*. IEEE, 2001, pp. 50–57.

[11] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 757–763, 1997.

[12] N. Bajaj, A. Routray, and S. Happy, "Dynamic model of facial expression recognition based on eigen-face approach," *CoRR*, 2013.

[13] T. Otsuka and J. Ohya, "Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences," in *Proceedings of the International Conference on Image Processing*, vol. 2. IEEE, 1997, pp. 546–549.

[14] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[15] H. Gu, Y. Zhang, and Q. Ji, "Task oriented facial behavior recognition with selective sensing," *Computer Vision and Image Understanding*, vol. 100, no. 3, pp. 385–415, 2005.

[16] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

[17] S. M. Platt and N. I. Badler, "Animating facial expressions," in *ACM SIGGRAPH Computer Graphics*, vol. 15, no. 3. ACM, 1981, pp. 245–252.

[18] K. Mase, "Recognition of facial expression from optical flow," *IEICE Transactions on Information and Systems*, vol. 74, pp. 3474–3483, 1991.

[19] Y. Yacoob and L. S. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 636–642, 1996.

[20] M. S. Bartlett, P. A. Viola, T. J. Sejnowski, B. A. Golomb, J. Larsen, J. C. Hager, and P. Ekman, "Classifying facial action," in *Advances in Neural Information Processing Systems*, 1996, pp. 823–829.

[21] E. Marg, "Descartes'error: Emotion, reason, and the human brain." *Optometry & Vision Science*, vol. 72, no. 11, pp. 847–848, 1995.

[22] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[23] M. Minsky, "Society of mind: a response to four reviews," *Artificial Intelligence*, vol. 48, no. 3, pp. 371–396, 1991.

[24] A. Pentland, "Looking at people: Sensing for ubiquitous and wearable computing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 107–119, 2000.

[25] A. van Dam, "Beyond wimp," *IEEE Computer Graphics and Applications*, vol. 20, no. 1, pp. 50–51, 2000.

[26] V. W. Zue and J. R. Glass, "Conversational interfaces: Advances and challenges," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1166–1180, 2000.

[27] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[28] C. Darwin, P. Ekman, and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[29] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of Personality and Social Psychology*, vol. 17, no. 2, p. 124, 1971.

[30] "Discussion: Emotion," https://ocw.mit.edu/courses/brain-and-cognitive-sciences/9-00sc-introduction-to-psychology-fall-2011/emotion-motivation/discussion-emotion/, (Date last accessed: Nov 20, 2016).

[31] "Facial expression analysis: The complete pocket guide," https://imotions.com/blog/facial-expression-analysis/, (Date last accessed: Nov 21, 2016).

[32] M. Suwa, N. Sugie, and K. Fujimora, "A preliminary note on pattern recognition of human emotional expression," in *International Joint Conference on Pattern recognition*, vol. 1978, 1978, pp. 408–410.

[33] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[34] J. Orozco, B. Martinez, and M. Pantic, "Empirical analysis of cascade deformable models for multi-view face detection," *Image and Vision Computing*, vol. 42, pp. 47–61, 2015.

[35] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image and Vision Computing*, vol. 32, no. 10, pp. 790–799, 2014.

[36] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[37] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[38] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. IEEE, 2005, pp. 886–893.

[40] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *International Conference on Image and Signal Processing*. Springer, 2008, pp. 236–243.

[41] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*. IEEE, 2011, pp. 314–321.

[42] 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*, vol. 29, no. 6, pp. 915–928, 2007.

[43] S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected sift features for 3d facial expression recognition," in *20th International Conference on Pattern Recognition (ICPR)*. IEEE, 2010, pp. 4125–4128.

[44] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 73–80.

[45] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 38–52, 2011.

[46] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *Image and Vision Computing*, vol. 24, no. 6, pp. 615–625, 2006.

[47] L. I. Smith, "A tutorial on principal components analysis," *Cornell University, USA*, vol. 51, p. 52, 2002.

[48] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," *Image and Vision Computing*, vol. 25, no. 12, pp. 1856–1863, 2007.

[49] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, 2007.

[50] M. Pantic and I. Patras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433–449, 2006.

[51] I. Mpiperis, S. Malassiotis, V. Petridis, and M. G. Strintzis, "3d facial expression recognition using swarm intelligence," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 2133–2136.

[52] P. Lemaire, M. Ardabilian, L. Chen, and M. Daoudi, "Fully automatic 3d facial expression recognition using differential mean curvature maps and histograms of oriented gradients," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–7.

[53] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013.

[54] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1940–1954, 2010.

[55] V. Le, H. Tang, and T. S. Huang, "Expression recognition from 3d dynamic faces using robust spatio-temporal shape features," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*. IEEE, 2011, pp. 414–421.

[56] B. Gong, Y. Wang, J. Liu, and X. Tang, "Automatic facial expression recognition on a single 3d face by exploring shape deformation," in *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, 2009, pp. 569–572.

[57] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1940–1954, 2010.

[58] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 200–205.

[59] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 65–72.

[60] M. F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proceedings of the 9th International Conference on Multimodal Interfaces*. ACM, 2007, pp. 38–45.

[61] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*. IEEE, 2011, pp. 298–305.

[62] U. Tariq, K.-H. Lin, Z. Li, X. Zhou, Z. Wang, V. Le, T. S. Huang, X. Lv, and T. X. Han, "Recognizing emotions from an ensemble of features," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1017–1026, 2012.

[63] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," in *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1998, pp. 454–459.

[64] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[65] Q. Zhao, D. Zhang, and H. Lu, "Supervised lle in ica space for facial expression recognition," in *International Conference on Neural Networks and Brain (ICNN)*, vol. 3. IEEE, 2005, pp. 1970–1975.

[66] R. Araujo and M. S. Kamel, "A semi-supervised temporal clustering method for facial emotion analysis," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 2014, pp. 1–6.

[67] T. Vandal, D. McDuff, and R. El Kaliouby, "Event detection: Ultra large-scale clustering of facial expressions," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–8.

[68] D. Y. Liliana, M. R. Widyanto, and T. Basaruddin, "Human emotion recognition based on active appearance model and semi-supervised fuzzy c-means," in *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2016, pp. 439–445.

[69] T. Senthilkumar, S. Rajalingam, S. Manimegalai, and V. G. Srinivasan, "Human facial emotion recognition through automatic clustering based morphological segmentation and shape/orientation feature analysis," in *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2016, pp. 1–5.

[70] X.-w. Chen and T. Huang, "Facial expression recognition: a clustering-based approach," *Pattern Recognition Letters*, vol. 24, no. 9, pp. 1295–1302, 2003.

[71] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, no. 1, pp. 160–187, 2003.

[72] Y. Zhu, L. C. De Silva, and C. C. Ko, "Using moment invariants and hmm in facial expression recognition," *Pattern Recognition Letters*, vol. 23, no. 1, pp. 83–91, 2002.

[73] M. F. Valstar, I. Patras, and M. Pantic, "Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data," in *Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*. IEEE, 2005, pp. 76–76.

[74] D. M. Tax, E. Hendriks, M. F. Valstar, and M. Pantic, "The detection of concept frames using clustering multi-instance learning," in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2010, pp. 2917–2920.

[75] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, and S. Z. Li, "Ensemble-based discriminant learning with boosting for face recognition," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 166–178, 2006.

[76] C. G. Kohler, T. Turner, N. M. Stolar, W. B. Bilker, C. M. Brensinger, R. E. Gur, and R. C. Gur, "Differences in facial expressions of four universal emotions," *Psychiatry Research*, vol. 128, no. 3, pp. 235–244, 2004.

[77] M. Osadchy, Y. L. Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *Journal of Machine Learning Research*, vol. 8, no. May, pp. 1197–1215, 2007.

[78] S. A. Sirohey, "Human face segmentation and identification," University of Maryland, College Park, Maryland, Tech. Rep. CAR-TR-695, 1998.

[79] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[80] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[81] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[82] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing*, vol. 16, no. 1, pp. 172–187, 2007.

[83] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 161–174, 2014.

[84] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Facial action recognition combining heterogeneous features via multikernel learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 993–1005, 2012.

[85] F. Tsalakanidou and S. Malassiotis, "Real-time 2d+ 3d facial action and expression recognition," *Pattern Recognition*, vol. 43, no. 5, pp. 1763–1775, 2010.

[86] S. Jaiswal, B. Martinez, and M. F. Valstar, "Learning to combine local models for facial action unit detection," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 6. IEEE, 2015, pp. 1–6.

[87] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[88] Y. Tang, "Deep learning using linear support vector machines," *CoRR*, 2013.

[89] S. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014, pp. 1431–1439.

[90] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *European Conference on Computer Vision*. Springer, 2012, pp. 808–822.

[91] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 443–449.

[92] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 503–510.

[93] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 467–474.

[94] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[95] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao, "Facial affect "in-the-wild": A survey and a new database," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.

[96] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[97] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition," in *Proceedings of the British Machine Vision Conference*, vol. 1, no. 3, 2015, p. 6.

[98] C. Goerick, D. Noll, and M. Werner, "Artificial neural networks in real-time car detection and tracking applications," *Pattern Recognition Letters*, vol. 17, no. 4, pp. 335–343, 1996.

[99] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[100] L. Ma and K. Khorasani, "Facial expression recognition using constructive feedforward neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 3, pp. 1588–1595, 2004.

[101] N. Bhattacharyya, R. Bandyopadhyay, M. Bhuyan, B. Tudu, D. Ghosh, and A. Jana, "Electronic nose for black tea classification and correlation of measurements with tea taster marks," *IEEE Transactions on Instrumentation and Measurement*, vol. 57, no. 7, pp. 1313–1321, 2008.

[102] R. Glasius, A. Komoda, and S. C. Gielen, "Neural network dynamics for path planning and obstacle avoidance," *Neural Networks*, vol. 8, no. 1, pp. 125–133, 1995.

[103] A. Cochocki and R. Unbehauen, *Neural networks for optimization and signal processing*. John Wiley & Sons, Inc., 1993.

[104] S.-C. Wang, "Artificial neural network," in *Interdisciplinary Computing in Java Programming*. Springer, 2003, pp. 81–100.

[105] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[106] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[107] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle *et al.*, "Greedy layer-wise training of deep networks," *Advances in Neural Information Processing Systems*, vol. 19, p. 153, 2007.

[108] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision*. Springer, 2014, pp. 818–833.

[109] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[110] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014.

[111] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[112] "Convolutional neural networks for visual recognition," http://cs231n.github.io/convolutional-networks/, (Date last accessed: Sep 14, 2017).

[113] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[114] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH*, vol. 2, 2010, p. 3.

[115] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 6645–6649.

[116] M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in neural machine translation," *CoRR*, 2014.

[117] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[118] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 609–616.

[119] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504 – 507, 2006.

[120] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8595–8598.

[121] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[122] "Exploring deep learning & CNNs," http://www.rsipvision.com/exploring-deep-learning/, (Date last accessed: Sep 20, 2017).

[123] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[124] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[125] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and Trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[126] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[127] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3626–3633.

[128] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[129] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[130] A. Ng, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, pp. 1–19, 2011.

[131] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[132] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[133] "Short history of the inception deep learning architecture," https://nicolovaligi.com/history-inception-deep-learning-architecture.html/, (Date last accessed: Oct 17, 2017).

[134] E. Alpaydin, *Introduction to machine learning*. MIT press, 2014.

[135] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, vol. 2, pp. 849–856, 2002.

[136] H. Liu, M. Shao, S. Li, and Y. Fu, "Infinite ensemble for image clustering," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1745–1754.

[137] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[138] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proceedings of the 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 2010, p. 65.

[139] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

Vita

Ahsan Jalal was born and raised in Rawalpindi, Pakistan. He completed his education at Sir Syed School and F.G Sir Syed College in Rawalpindi. He earned a bachelors degree in Electrical Engineering from National University of Sciences and Technology (NUST), Islamabad, Pakistan. Upon his graduation in 2014, Ahsan Joined the Center of Excellence in FPGA and ASIC (CEFAR) to work as a research assistant on various projects related to image processing and Android applications for smart security. He was also a research assistant at TUKL-NUST Research Center where he worked on several machine learning projects and research works. He published his research on automatic fish detection, tracking and classification system under unconstrained environment in top-tier journal "Limnology and Oceanography". In 2016, Ahsan joined master's program in Electrical Engineering at the American University of Sharjah (AUS), UAE. He also worked as a graduate research and teaching assistant from February 2016 onwards. During his work and study at AUS, he published a conference paper titled "The LFW-gender Dataset" in ACCV-2016. His forthcoming publication is titled "Semi-supervised Clustering of Unknown Expressions". His research interests include image processing, computer vision and machine learning.