FACE IDENTIFICATION AND CLUSTERING

BY ATUL DHINGRA

A thesis submitted to the
Graduate School—New Brunswick
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Master of Science
Graduate Program in Computer Science

Written under the direction of
Dr. Vishal Patel, Dr. Ahmed Elgammal
and approved by

_____

_____

_____

New Brunswick, New Jersey

May, 2017

ABSTRACT OF THE THESIS

Face Identification and Clustering

by Atul Dhingra

Thesis Director: Dr. Vishal Patel, Dr. Ahmed Elgammal

In this thesis, we study two problems based on clustering algorithms. In the first problem, we study the role of visual attributes using an agglomerative clustering algorithm to whittle down the search area where the number of classes is high to improve the performance of clustering. We observe that as we add more attributes, the clustering performance increases overall. In the second problem, we study the role of clustering in aggregating templates in a 1:N open set protocol using multi-shot video as a probe. We observe that by increasing the number of clusters, the performance increases with respect to the baseline and reaches a peak, after which increasing the number of clusters causes the performance to degrade. Experiments are conducted using recently introduced unconstrained IARPA Janus IJB-A, CS2, and CS3 face recognition datasets.

Acknowledgements

I would like to first thank my thesis advisor, Dr. Vishal Patel who provided gave me inspiration and encouragement throughout. I would also like to thank my thesis co-advisor, Dr. Ahmed Elgammal who helped me navigate through this journey. I would also like to thank all my lab-mates for the exchange of ideas, academic and otherwise. Finally, I express my profound gratitude for my family who has helped me arrive at this point in my academic career.

Dedication

I dedicate this thesis to my family.

Table of Contents

Chapter 1

Face Recognition

1.1 Introduction

Face recognition has been actively studied over the past few decades which has led to satisfactory performances in recognition rates in controlled scenarios. But, in an unconstrained environment, face recognition is still a hard problem. A number of datasets have been thus developed to study face recognition in these scenarios that include LFW [9], PubFig[2] and IJBA[12]. The intuitive pipeline[1] is shown in figure 1.1 for face recognition, that includes face detection and tracking, face alignment, feature extraction and matching, described in sections below.

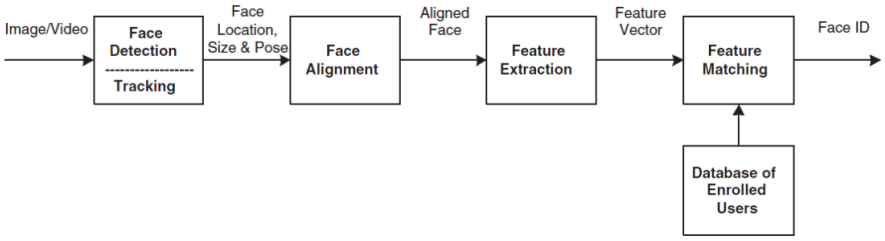

Figure 1.1: Face Recognition Pipeline

1.1.1 Acquisition

There are a few challenges that hinder the progress of face recognition in an unconstrained environment, which include challenges such as pose, illumination, and expression (PIE). Some of the other notable challenges include aging, cosmetics and resolution of the image. A lot of datasets have been developed that provide challenging

media(images, videos, templates) so that algorithms can be developed to deal with these issues. Yale and YaleB[27] was introduced in 1997 that highlighted the challenges in illumination conditions, AR dataset[28] in 1998 highlighted occlusion apart from different emotions, and illuminations. Some of the more notable datasets in recent past are LFW[9] and PubFig[2] that contain huge amounts of images and deal with the face representation in the wild. One of the most challenging datasets as of now is the recently introduced unconstrained IARPA Janus IJB-A, CS2 and CS3 face recognition datasets[12].

1.1.2 Normalization and Alignment

Some of the pose and illumination artifacts are removed by normalization. There has been a lot of work that deals with this task. Depending on the applications, the issue of normalization can either be handled during the acquisition phase, where during the collection of the database, the acquisition parameters, such as capture device, ambient light are fixed. But, in the case where we want to develop algorithms invariant to these artifacts in unconstrained settings, learning from data in such preferential environment is averse to learning in the real-world settings. In such a case, post-processing of the collected data is done.

Illumination Normalization

We can handle illumination normalization during the acquisition phase, by making sure that the illumination remains the same throughout. As some of the datasets are collected in the real world settings, the natural illumination affects the final dataset. In such a case active devices such as thermal infra-red images, near infra-red images etc. can be used that provides its own light source to illuminate the object.

In case this is not possible, such as images in the wild, normalization is done during post-processing to generate illumination invariant features. This can be done by using methods such as linear subspace, illumination cone, generalized photometric stereo, photometric normalization, reflectance model [3] etc. Some of the most studied models include, Self Quotient Images [17],Logarithmic Total Variation [18], Gradient Faces[19],

Robust Albedo Estimation[20].

Pose Normalization

The images captured during the data acquisition phase can be constrained such that the pose of the captured images are consistent. But, is not the best solution, as even a slight error in capturing would result in a completely different image vector. Therefore, in such a case pose normalization is done during post-processing. The approach, in this case, is to find landmarks in the image that would remain consistent throughout, no matter how much shifted the image is. Some of these landmarks include the eyes, the nose, and the lips. Once these landmarks are detected, the image can be normalized based on these set of points. One such method that takes into account such an approach is called Geometric warping[35] where in-place pose normalization can be achieved. But, this approach cannot help in the case where there is an out-of plane rotation, for a case more robust methods are required. This follows from the fact that in an out-of plane rotation, pitch, roll, and yaw all have to be normalized. Some of the more used methods studied are, Incremental face alignment[23], Deep Face Alignment[21],Face Frontalization[22].

1.1.3 Features/Recognition

Features are distinct and unique properties of an entity, that can be used to distinguish it from others. These features are important as they form the framework for recognition of these entities. In a face recognition system, facial features could include, the shape of a person's face, eye color, the distance between eyes, etc. These features could either be hand-crafted, or they could be learned features.

Hand-Crafted features

Hand-crafted features as the name suggests is created manually by observing uniqueness in some aspect of the object. At the lowest levels, edges, lines, and corners form features, in a complex object, such as a face, a combination of these low-level features by hand is known as hand-crafted features. There are a few hand-crafted features that have

been used extensively, such as SIFT[32], HOG[33], LBP[34], etc. In such a framework, a classifier is trained using the hand-crafted features.The classification/recognition can be done using SVM[36] , SRC[37] and Subspace methods such as PCA[38], LDA [39] etc.

Learned Features

Instead of coding the features by hand, features can also be learned from the data. This ensures an optimal representation given the data. At the end, a simple classifier can be used for classification. There are a few methods that are used in such a scenario, which include Dictionary Learning, Neural Networks etc.

1.2 Protocols

Recognition is a term with wide scope when it comes to Face biometrics, as it encompasses a lot of authentication protocols, there are a few widely used authentication types that have been described below.

1.2.1 Identification

In an identification problem, the question asked is, whether a given person exists in our system or not. The output from such a system is either Identified or Not-identified depending on whether that person exists in the given database.

1.2.2 Verification

In a verification protocol, given an instance of a user, we check if it matches the sample of the same user in our system. The output in such a scenario is a similarity score which defines how closely the new sample matches to the one already in the system

1.2.3 Search

In a search scenario, given a query image, we need to find all the instances of that person in the database. The output, in this case, top-k hits of the subject

1.3 Metrics

1.3.1 Error Statistics

A few of the more commonly used error statistics are False Match(Type I Error), False Non-Match (Type II Error), True-positive Identification Rate(TPIR), False-positive Identification-error Rate (FPIR).

True-positive Identification Rate(TPIR) The True-positive Identification Rate (TPIR) is the proportion of identifications by enrolled subjects in which the subjects correct class is returned. [24]

False-positive Identification-error Rate (FPIR) The False-positive Identification-error Rate (FPIR) is the proportion of identifications by users not present in the system, which is returned. FPIR cannot be computed in closed-set identification, as all users are enrolled in the system [24]

1.3.2 Decisions

1.3.3 Metric curves

There are a few metric curves that are used to plot the decisions, that include Reciever Operating Characteristics(ROC), Detection Error Tradeoff(DET), Cumulative Match Curve(CMC).

CMC A CMC curve plots the Probability of identification versus the Rank as shown in figure 1.2

1.3.4 Result interpretation

The result interpretation depends on the type of face recognition application. Some of the more used interpretations include Accuracy, Precision and Recall and F-Measure.

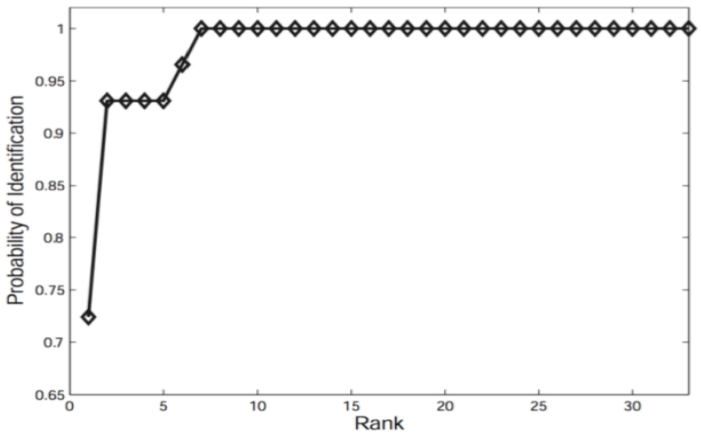

Figure 1.2: CMC Curve[1]

F-measure The F-measure is given in equation 1.1 where P is Precision and R is Recall

$$F_\beta = \frac{(\beta^2 + 1)P.R}{\beta^2 P + R} \tag{1.1}$$

The F-1 measure is widely used where $\beta = 1$, such that F-1 measure is the harmonic mean between precision and recall. The value of F-measure, therefore, is always between 0 and 1, and the higher the value, better is the performance of the recognition algorithm.

Precision and Recall Precision is defined as the ratio of True positives(TP) to the sum of True positives and False Negatives(FN) as shown in figure 1.3

$$Precision = \frac{TP}{TP + FP} \tag{1.2}$$

Recall is similarly defined as the ratio of true positives over the sum of true positives and false positives(FP) as in figure 1.2

$$Recall = \frac{TP}{TP + FN} \tag{1.3}$$

Chapter 2

Face Clustering

2.1 Introduction

Clustering is an unsupervised classification of patterns such as data items, feature vectors, or observations. In such a setting, given unlabelled data points, we have to group them based on a metric(ℓ_2,ℓ_p, Mahalanobis etc.). Clustering is a difficult problem, as we need to know a priori about the number of clusters or the stopping criterion. Clustering has a lot of applications such as exploration, segmentation in cases where the prior information about the data is not available. The pipeline[5] for clustering is given in figure 2.1

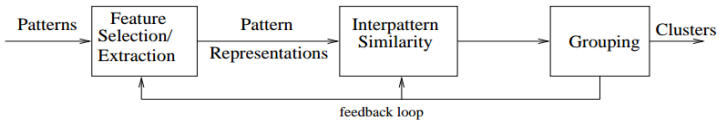

Figure 2.1: Pipeline for Clustering

A good representation the given data points/patterns is achieved by feature extraction. Once these features are computed, the clusters are merged/divided based on the inter-pattern similarity and the type of clustering. This process goes on until a stopping criterion such as a distance threshold or max number of clusters is met.

2.2 Clustering Techniques

2.2.1 Hierarchical

Hierarchical clustering seeks to build a hierarchy of clusters such that it yields a dendrogram that represents the nested grouping of patterns and similarity levels[5]. These fall into two categories, agglomerative clustering, and divisive clustering.

Agglomerative

This is a bottom-up approach where each observation starts as an independent cluster, and pairs of clusters are merged based on the hierarchy and a stopping criterion. The merging of the clusters is based on certain linkage criterion, such as Single Link where the minimum distance between the points is used to merge the cluster. In the case of complete-link clustering, the clusters are merged based on the maximum distance between the data points of the two clusters. There are other order statistics that are used such as mean, centroid, group average, etc. to perform these linkages as well.

Divisive

In a divisive clustering framework, a top-down approach is followed such that all the data points start out in a single cluster, and they are split into different clusters moving down the hierarchy.

2.2.2 Partitional

In the case where construction of dendrograms is computationally inefficient/impossible, partitional methods are employed where a single partition of the data is obtained instead of a structure. The issue with using partitional clustering techniques is the fact that we need to know a priori the number of clusters/ partitions we need to perform. Partitional clustering is produced by optimizing a criterion function defined either locally or globally[5]. Some of the most common criterion used are squared error method as represented in equation 2.1 [5], where X is the patterns set of the clustering L, which

contains K clusters, such that $x_i^{(j)}$ is the i^{th} pattern belonging to the j^{th} cluster and c_j is the centroid of the j^{th} cluster

$$e^2(X, L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} ||x_i^{(j)} - c_j||^2 \tag{2.1}$$

A widely used method that uses squared error criterion is the k-means algorithm, where k points are randomly picked as the centroid and the cluster's center are recomputed until convergence by assigning each point to the cluster with the closest centroid.

2.3 Evaluation

The ultimate aim for clustering algorithm is to attain high intra-cluster similarity and low inter-cluster similarity. There are few evaluation metrics that are widely used to access the quality of the clustering. Some of these are Purity, Precision, and Recall, F-measure and compactness[29]. In our work, we use pair-wise precision and recall as defined in [6]

Pairwise Precision is the same class fraction of pairs of data points within a cluster over the total number of same cluster pairs within the dataset. [6].

Pairwise Recall is the fraction of within class pairs of data points, that are placed in the same cluster, over the total number of same-class pairs in different clusters. [6].

2.4 Recent Works and Motivation

In the problem of clustering faces, given unlabelled face images, we need to divide them into clusters, using a good feature space representation and a distance metric as shown in figure 2.2[6].

There has been a lot of work in the area of face clustering that tries to improve the clustering accuracy. Zhu et al.[7] came up with Rank-Order Distance that is robust to both noise and outliers and can handle non-uniform cluster distribution like varying densities, shapes, and sizes of clusters. It calculates the dissimilarity between two faces based on their neighbouring information using ℓ_1 distance motivated by the fact that the same person shares top neighbours. The sub-clusters formed due to variation in

Figure 2.2: Pipeline for Face Clustering [6]

pose illumination and expression, are subsequently merged agglomeratively using rank-order distance using a certain threshold or cluster level rank order distance to avoid the problem of too many high-precision, tight sub-clusters in the case of just using rank-order distance.

Otto et al[6] used the same idea as Zhu et al. [7] on a larger scale, and therefore modified the algorithm to work on a large data setting. The effective and efficient Rank-order clustering algorithm used k-d tree algorithms to compute a small list of nearest neighbour, as the input size of data in order of millions, generating all the neighbours, as in the case of Zhu et al. [7] would be computationally hard. It used a single linkage agglomerative clustering algorithm based on a threshold to further compute the clusters and uses a pairwise F-measure to report the results on LFW dataset[9].

Zhu et al[8] came up with an algorithm to iteratively merge high precision clusters based on heterogeneous context information such as common-scene, people co-occurrence, human attributes and clothing information, such that the resulting clusters also have high recall.

Clustering is hard as the performance decreases as the number of classes increases as it is evident in figure 2.3. Therefore our work is motivated by this challenge to whittle down the search domain in clustering using visual attributes to improve the clustering accuracy.

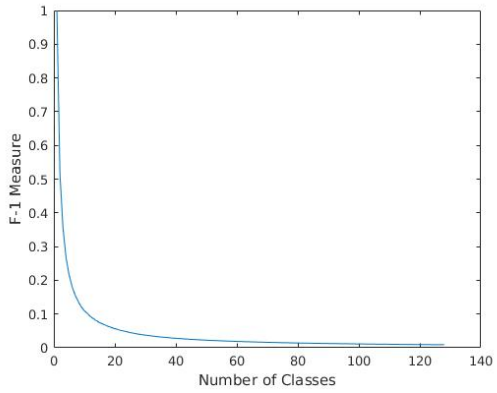

Figure 2.3: F-measure plot as the number of classes increases

2.5 Experiment

In our work, media averaging is done on CNN features that are computed from the IJBA CS2[12] samples in order to obtain templates. As the CS2[12] follows a template to template matching protocols, we perform clustering on these media averaged templates. Media averaging is shown in figure 2.4, where the video frames with the same media ID are averaged, and the resultant is then averaged with the images that belong to the same template ID.

Figure 2.4: Media Average template [12]

Once these templates are obtained, we use agglomerative clustering as defined in section 2.2.1 where each template starts out as a different cluster are merged based on the stopping criterion of max number of clusters, as we have prior information of classes from the dataset. We use the average linkage with the cosine metric for merging these clusters based on the inter-pattern similarity. The templates are further divided

into disjoint sets based on ground truth attributes from CS2 [12]. The template is divided into a Male subset, a female subset. The male subset is further divided into two different disjoint subsets based on the skin color attribute.The accuracy of the algorithm is reported based on pairwise F-1 score described in section 2.3

2.6 Results

The algorithm is evaluated on IJBA CS2 dataset [12] that contains 500 subjects with 5,397 images and 2,042 videos split into 20,412 frames. The IJBA CS2 evaluation protocol consists of 10 random splits that contain 167 gallery templates and 1763 probe templates. The algorithm is evaluated on these 10 splits on JC's[25] and Swami's[26] deep features. The evaluated results on Swami's [26] features are shown in table 2.1 and figure 2.6. The evaluated results on JC's[25] features are shown in table 2.2 and figure 2.6

Split	Base	Male	Female	M+Skin 1	M+Skin 3
1	0.7281	0.7349	0.7542	0.7417	0.82
2	0.7134	0.6756	0.8364	0.7384	0.7728
3	0.6817	0.7025	0.7001	0.74	0.7336
4	0.7349	0.7309	0.7676	0.7633	0.7683
5	0.6066	0.6133	0.6418	0.6517	0.648
6	0.6756	0.6729	0.7577	0.7213	0.7524
7	0.7309	0.7651	0.7294	0.7513	0.816
8	0.6561	0.6875	0.616	0.7648	0.8001
9	0.6531	0.6845	0.7939	0.7391	0.7095
10	0.6645	0.6907	0.6875	0.7296	0.7504
Average	**0.68449**	**0.69579**	**0.72846**	**0.73412**	**0.75711**

Table 2.1: Evaluation of algorithm on Swami Features[26]

2.7 Conclusion

We observe in table 2.1 and table 2.2 that as we use more attributes, the clustering result improves. We can, therefore, assert that by using visual attributes we are narrowing down the search domain of the algorithm to boost the performance of clustering.

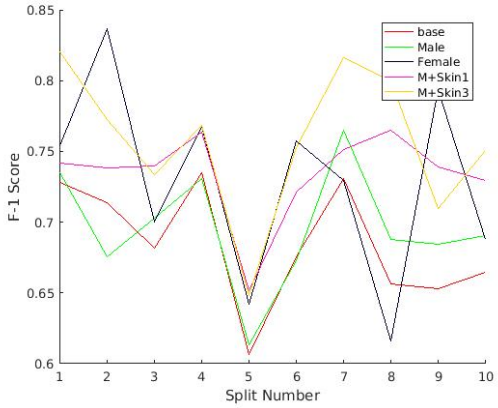

Figure 2.5: F-1 score of the experiment using Swami[26] features

Split	Base	Male	Female	M+Skin 1	M+Skin 3
1	0.6847	0.6854	0.7591	0.6807	0.7424
2	0.682	0.6512	0.753	0.6334	0.728
3	0.6356	0.6675	0.6673	0.7044	0.6997
4	0.6716	0.6597	0.7197	0.7111	0.7325
5	0.5658	0.5706	0.6036	0.5999	0.6083
6	0.6633	0.6385	0.7586	0.651	0.738
7	0.6832	0.6931	0.6844	0.7266	0.8204
8	0.6534	0.706	0.5862	0.7242	0.7833
9	0.6157	0.6563	0.6405	0.7109	0.6677
10	0.6663	0.6783	0.6712	0.7133	0.7694
Average	**0.65216**	**0.66066**	**0.68436**	**0.68555**	**0.72897**

Table 2.2: Evaluation of algorithm on JC Features[25]

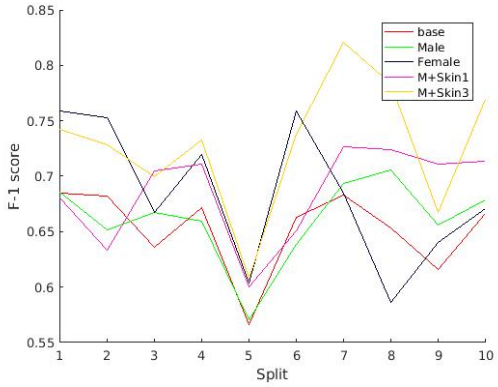

Figure 2.6: F-1 score of the experiment using JC[25] features

Chapter 3

Video Based face tracking and Identification

3.1 Introduction

In this work, we focus on a face identification task where the target is a multi-shot video and is annotated only once in one of the frames, and we need to search the annotated subject in a given gallery of images. The advantage over image to image retrieval in this case is that with a probe video, we have a lot more information and exemplars of the subject of interest and we can leverage this information to come up with a more robust representation that is invariant to the PIE challenges in face recognition.

Traditionally, for a video to image retrieval task, the probe video is single shot where frame by frame bounding box of the subject of interest is provided as in the case of Youtube Faces [30]. For our work, we study an open set 1:N protocol using full motion video as probe where the probe video is multi-shot. In this setting, the subject of interest is annotated only for one of the frames, and the subject may or may not reappear in the subsequent shots. Therefore, matching a subject of interest from multi-shot video to gallery is a difficult task as we cannot use the traditional methods of a frame by frame bounding box tracking for the target face, because tracking algorithms are prone to drifting.

A baseline approach to this problem is just to use the initial representation of the user annotated face of the subject to search for the subject in the gallery. But, the initial representation may not always be full frontal and devoid of any pose, illumination and expression variations. Hence, finding the subsequent appearance of the subject in the video is required to come up with a very robust representation of the subject. This is relatively easy in a single-shot video, where the entire video is a single shot, and there is no break in continuity. This can be achieved by making use of the temporal information

and tracking the subject throughout in the video. But, in the case of multi-shot video, this task is relatively hard in a multi-shot video.

3.2 Motivation and Recent Works

The problem of face recognition as described in section 1.2, can be looked at in the terms of face verification and face identification. In face verification protocol one-to-one similarity is computed between the probe and the reference image. In face identification, on the other hand, one-to-many similarity between the probe and gallery is computed. With LFW[9] the face there were attempts to solve the face identification in the case where the dataset was unconstrained. Even so, there was a near-frontal selection bias while constructing the LFW[10], hence the results are not representative of the set containing large in-the-wild pose variation. Also, because recent studies, [11] suggest the algorithmic performance of Face recognition algorithms is sub-par to humans, performance on unconstrained datasets with extreme pose, illumination, and expression are still lagging. One such challenging dataset is IJBA[12] that provides protocols for template-based verification and identification. The dataset consists of images and videos of subjects that are manually annotated and the performance evaluation is over a template, such that set of all media is combined into a single representation. Generating a robust representation in the form of a template is of utmost importance due to the large variation in pose, illumination, and expression. In our work, we improve an existing algorithm by template aggregation using clustering.

There has been some work on templates and multi-shot video to gallery retrieval that has motivated our work in this direction. N. Crosswhite et al. [10] presented template adaptation, a type of transfer learning that works on the IJBA dataset [12] on one-to-many face identification protocol using CNN features, and a template specific one-vs-rest linear SVM. In their work, they learned a transfer learning mapping such that the source domain is the CNN features learned, and the target domain is the template of new subjects. This work uses encoding from the penultimate layer of VGG-Face[13] using an anisotropically scaled face crop of 224x223x3, followed by learning

an L2-regularized L2-loss primal SVM with class weighted hinge loss objective[10] as expressed in equation 3.1.

$$min_w \frac{1}{2} w^T w + C_p \sum_{i=1}^{N_p} max[0, 1 - y_i w^T x_i]^2 + C_n \sum_{j=1}^{N_n} max[0, 1 - y_j w^T x_j]^2 \tag{3.1}$$

such that C_p is the regularization constant for N_p positive observations obtained via average media encodings in the template, and C_n for negative observation obtained via large external negative features.

Ching Hui et al. [14] combine the work of Template Adaptation [10] and context-assisted clustering [8] to propose a Target Face Association(TFA) technique [14] that retrieves a set of representative face images from multi-shot video that is likely to have the same identity as the target face which is then used as a robust representation based on which the subject is looked up in a gallery of images. An OTS tracking technique[15] is used to track the target face. These images are treated as the initial positive training set(S_p). The faces are pre-associated [14] by selecting highest Intersection over Union(IOU)[16] of face detection bounding box with the with tracking bounding boxes for the first k-frames. Ching Hui et al. learns a target specific linear SVM iteratively from pre-associated face images(positive samples) from the target video and negative samples(S_n) obtained from the cannot-link constraint[8]. In the cases where the target video cannot establish cannot-link constraints, due non-existence, an external dataset (S_b) is prepared for negative instances of the SVM. Their work uses two different models, wherein model one, the linear SVM is solved using the max-margin framework, where the training data is the union of all the three sets, i.e $\{(x_i, y_i)|i \in (S_p \cup S_n \cup S_b)\}$ are used for training. In model 2, the set S_b is used only when there are no within-video negative instances. The robust face representation[14] is given in equation 3.2 ,

$$x^{fa} = \frac{1}{|A|} \sum_{i \in A} x_i \tag{3.2}$$

3.3 Method

Once the TFA [14] algorithm outputs the positive samples from the SVM, it simply averages these features as shown in equation 3.2 to obtain the robust representation.

In our work, however (TFA-C), we leverage a clustering algorithm to aggregate the features at the end of TFA. We use an Approximate Nearest Neighbour k-means++ using VLFeat library [31] algorithm such that the k data points that are picked greedily are maximally different. It is optimized using Approximate Nearest Neighbour algorithm that uses a randomized k-d tree. The max number of comparisons is limited to 100 and the number of trees is limited to 2 to trade off between speed and accuracy. The clustering is done by varying the number of clusters between 1 and 20. In the case where the number of samples is less than the number of clusters, the maximum cluster value is clipped to the maximum number of samples.

3.4 Results

JANUS CS3 is an extended version of IJBA dataset [12] that contains 11,876 images and 7245 video clips of 1870 subjects. CS3 provides 11 different protocols, that include Identification, Verification and clustering tasks. In our work, we focus on Protocol 6, i.e CS3 1:N Multi Open Set (Video). In Protocol 6 there are 7195 probe templates, where each template is evaluated with respect to two disjoint galleries. There are 940 and 930 templates in Gallery 1 and Gallery 2, respectively. In this case given a video and the annotation of the subject of interest in the first frame, we need to search for a mated template in the gallery for a given probe template. As protocol 6 is an open-set identification problem, there exist some probe templates for which there are no mated templates in the gallery. Therefore, the ranking accuracy is evaluated only for those probe templates that have a mated template in the gallery, demonstrating the closed-set search. For these 20 clusters, the Rank-1, Rank-5, Rank-10, Rank-25, TPIR results are plotted for both JC[25] and Swami[26] features. These results are shown in figure 3.1 to figure 3.6

On an average k=7 clusters work best in respect to Rank-k accuracy and TPIR rate. The computed results for k=7 for JC[25] are given in table 3.1 and the output on Swami's[26] features are given in table 3.2. As Ching Hui et al.[14] report their results on the average of these two features, we also report the average output in table 3.3.

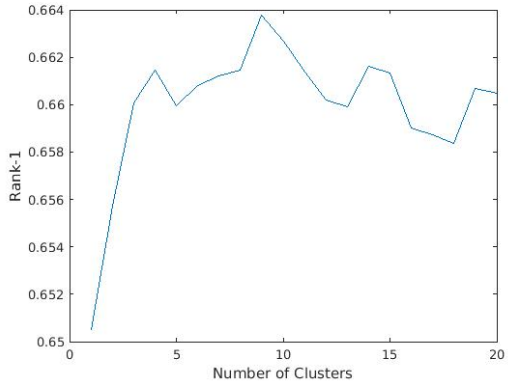

Figure 3.1: Rank-1 CMC Plot

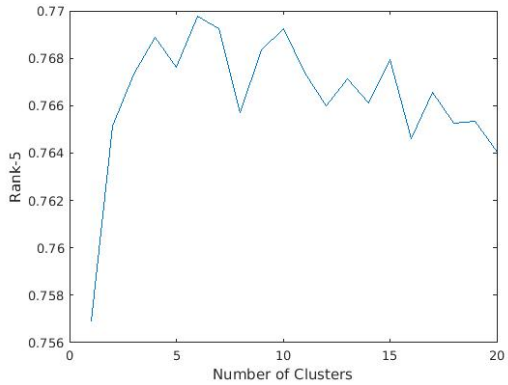

Figure 3.2: Rank-5 CMC Plot

As we can clearly see, the results for TFA-C in table 3.3 is better than the original TFA algorithm in table 3.4 we can state that TFA-C performs better than TFA[14]

3.5 Conclusion

We observe that as the number of clusters(k) are increased for the template aggregation, the identification rate increases to a point and deprecates after that. Based on the averages, we observe cluster numbers, k=7 works the best for identification rate in closed set search as shown by the CMC Rank curves and also in the open set search as shown by the CMC TPIR curves. We conclude that our method TFA-C outperforms the existing TFA algorithm by a significant margin.

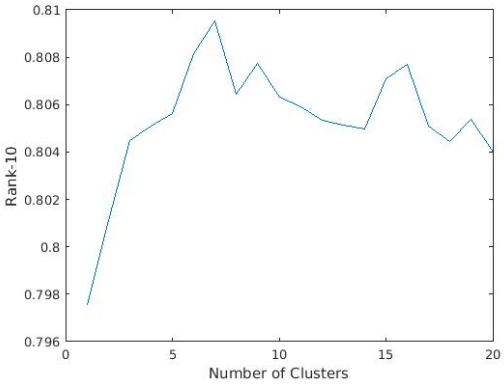

Figure 3.3: Rank-10 CMC Plot

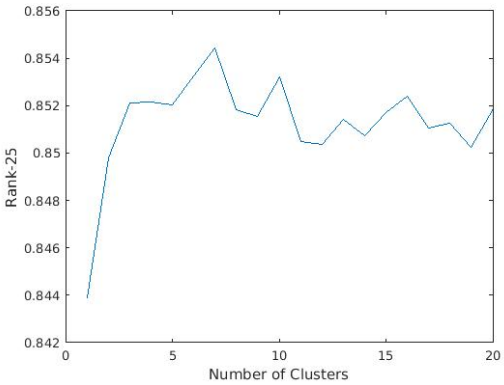

Figure 3.4: Rank-25 CMC Plot

3.6 Acknowledgement

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon

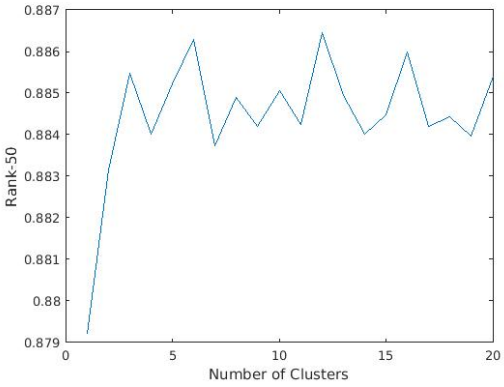

Figure 3.5: Rank-50 CMC Plot

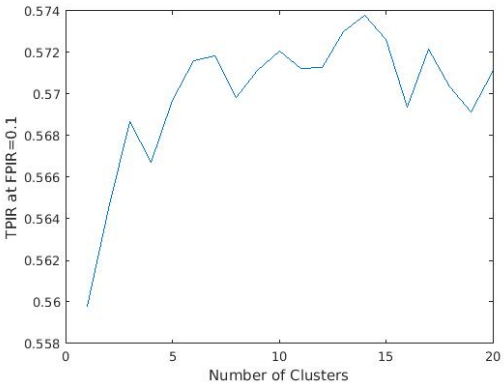

Figure 3.6: CMC plot of TPIR at FPIR=0.1

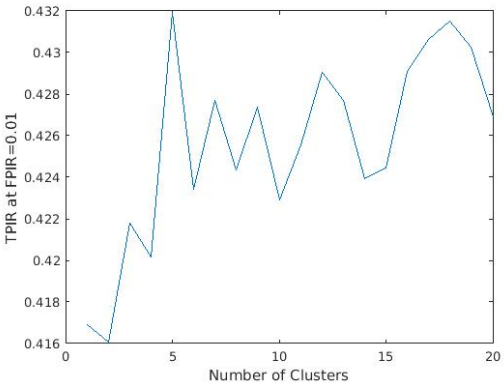

Figure 3.7: CMC plot TPIR at FPIR=0.01

	Rank-1	Rank-5	Rank-10	Rank-25	Rank-50	TPIR at FPIR=0.1	TPIR at FPIR=0.01
Gallery 1	0.68452	0.7913	0.83165	0.87339	0.90226	0.59757	0.45878
Gallery 2	0.70122	0.81287	0.84348	0.88209	0.90678	0.61496	0.47583
Average	0.72383	0.82991	0.85983	0.89843	0.9193	0.64765	0.50678

Table 3.1: Evaluation of TFA-C using JC features[25] on k=7 clusters

	Rank-1	Rank-5	Rank-10	Rank-25	Rank-50	TPIR at FPIR=0.1	TPIR at FPIR=0.01
Gallery 1	0.55	0.66782	0.71644	0.77338	0.81806	0.46157	0.32963
Gallery 2	0.57662	0.68472	0.7338	0.79097	0.83449	0.46968	0.29097
Average	0.59861	0.70856	0.75926	0.81042	0.84815	0.49606	0.34861

Table 3.2: Evaluation of TFA-C using Swami features[26] on k=7 clusters

	Rank-1	Rank-5	Rank-10	Rank-25	Rank-50	TPIR at FPIR=0.1	TPIR at FPIR=0.01
Gallery 1	0.61726	0.72956	0.77404	0.82339	0.86016	0.52957	0.39421
Gallery 2	0.63892	0.7488	0.78864	0.83653	0.87064	0.54232	0.3834
Average	0.66122	0.76924	0.80954	0.85443	0.88373	0.57186	0.4277

Table 3.3: Evaluation of TFA-C using average both features on k=7 clusters

	Rank-1	Rank-5	Rank-10	Rank-25	Rank-50	TPIR at FPIR=0.1	TPIR at FPIR=0.01
Gallery 1	0.6689	0.7875	0.8264	0.8803	0.913	0.5701	0.3892
Gallery 2	0.5514	0.6803	0.7315	0.7926	0.8394	0.4245	0.2931
Average	0.6101	0.7339	0.779	0.8365	0.8762	0.4973	0.3411

Table 3.4: TFA[14] results on average of both features on k=7 clusters

Chapter 4

Appendix

Photo-Sketch

Facial sketches are an essential part of forensics in law enforcement, particularly in those cases where the only evidence is in the form of eye-witness testimony. Facial sketches are of two types, Forensic Sketches that are drawn by forensic artists, and Composite Sketches that are created using computer software [4]. Once the sketches are drawn from either of these methods, it allows the law enforcement to apprehend the person of interest based on it. Several works have tried to automate this process by automatically matching[4] the sketches to the criminal database. Figure 4.1 shows composite and forensic sketches corresponding to the mugshot images as developed by Klum et al.[4]. They also show that the matching accuracy of composite sketches is higher than that of the forensic sketches. As evident from the figure 4.1, composite sketches are more close to the mugshot images in the domain, and hence they have a better matching accuracy.

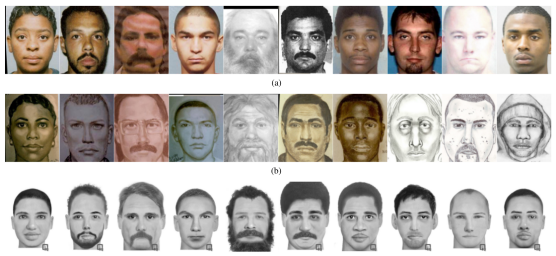

Figure 4.1: Forensic and Composite sketched corresponding to mugshot images [4]

Motivated by the fact, that at the end the ultimate aim of sketches is matching, we wanted to develop automatic sketches in the mugshot domain. For our work, we used the PubFig dataset to develop single average template faces for the attributes using

one attribute and two attributes as shown in figure 4.2 and figure 4.3

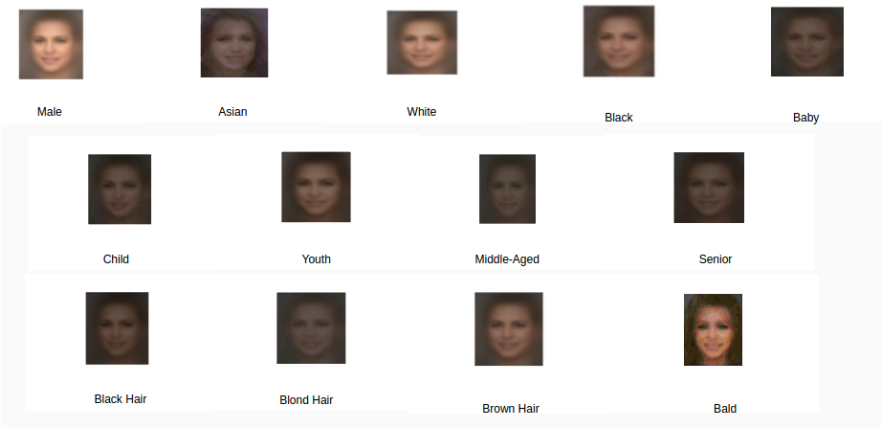

Figure 4.2: Average faces using one attribute

Figure 4.3: Average faces using two attributes

As evident from figure 4.2 and figure 4.3 the average template suffer high illumination artifacts and there is a bias across not only the subjects but across the attributes. So a trade-off needs to met so that the dataset is balanced not only in the subjects but also, attributes. Due to the lack of such a curated dataset and the ill-posed problem, we will like to work further on this problem by either developing a dataset in the future, or utilizing a dataset if any is created that balances classes across not only subjects, but attributes as well.

Vita

The author of my thesis

2017 M.S in Computer Science, Rutgers University, USA

2014 B.E in Instrumentation and Control, University of Delhi, India

2016-2017 Graduate assistant, Department of Computer Science, Rutgers University

2015-2016 Teaching assistant, Department of Computer Science, Rutgers University

2015-2016 Grader, Department of Computer Science, Rutgers University

2011-2015 Visiting Researcher, IIT-Delhi, India

References

[1] S.Li and A.Jain, (ed). *Handbook of Face Recognition*, Springer-Verlag, 2005

[2] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar, *Attribute and Simile Classifiers for Face Verification*, International Conference on Computer Vision (ICCV), 2009.

[3] Xuan Zou, Kittler, J. , Messer, K , *Illumination Invariant Face Recognition: A Survey*, 2007

[4] S. Klum, H. Han, A. K. Jain, and B. Klare *Sketch based face recognition: Forensic vs. composite sketches*, In Proc. ICB, 2013

[5] A.K. Jain, M.N. Murty, P.J. Flynn, *Data Clustering: A review* , 1999

[6] Otto,Wang,Jain, *Clustering Millions of Faces by Identity*,2016

[7] Chunhui Zhu, Fang Wen, Jian Sun, *A Rank-Order Distance based Clustering Algorithm for Face Tagging*, In: CVPR. (2011)

[8] L. Zhang, D. V. Kalashnikov, and S. Mehrotra,*A unified framework for context assisted face clustering* ,In ICMR, pp. 916, 2013

[9] G. Huang, M. Ramesh, T. Berg, and E. Learned Miller, *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*, Technical Report 07-49, UMass, 2007.

[10] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman, *Template adaptation for face verification and identification*, arXiv preprint arXiv:1603.03958, 2016.

[11] Phillips, J., Hill, M., Swindle, J., OToole, A., *Human and algorithm performance on the pasc face recognition challenge*, In: BTAS. (2015)

[12] Navaneeth Bodla, Jingxiao Zheng, Hongyu Xu, Jun-Cheng Chen, Carlos Castillo, Rama Chellappa, *Deep Heterogeneous Feature Fusion for Template-Based Face Recognition*, arXiv preprint arXiv:1702.04471, 2017

[13] Parkhi, O., Vedaldi, A., Zisserman, A. *Deep face recognition*, In: BMVC. (2015)

[14] Ching Hui, *Video-Based Face Association and Identification* , In: FG 2017

[15] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, *High-speed tracking with kernelized correlation filters*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3):583596, Mar. 2015

[16] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisser- man, *The Pascal visual object classes (VOC) challenge* ,International Journal of Computer Vision, 88(2):303338, Sep. 2009.

[17] Haitao Wang, Stan Z. Li, Yangsheng Wang, Jianjun Zhang,*Self quotient image for face recognition* ,In :IEEE FG 2004

[18] T. Chen, W. Yin, X. S. Zhou, D. Comaniciu, and T. Huang, *Total variation models for variable lighting face recognition*, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), 28 (2006), pp. 15191524.

[19] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu, *Face recognition under varying illumination using gradientfaces*, IEEE Trans. Image Process., vol. 18, no. 11, pp. 25992606, Nov. 2009

[20] S. Biswas, G. Aggarwal, and R. Chellappa, *Robust estimation of albedo for illumination-invariant matching and shape recovery*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 2, pp. 884899, Mar. 2009.

[21] Taigman, Y., Yang, M., Ranzato, M. and Wolf, L. *Deepface: closing the gap to human-level performance in face verification*, In Proc. Conference on Computer Vision and Pattern Recognition 17011708 (2014)

[22] T. Hassner, S. Harel, E. Paz, and R. Enbar. *Effective face frontalization in unconstrained images*, Proc. Conf. Comput. Vision Pattern Recognition, 2015.

[23] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, *Incremental face alignment in the wild*, In Proceedings of IEEE Intl Conf. on Computer Vision & Pattern Recognition (CVPR 2014)

[24] *Biometrics Metric Report*, `http://www.usma.edu/ietd/docs/BiometricsMetricsReport.pdf`

[25] J.-C. Chen, V. M. Patel, and R. Chellappa, *Unconstrained face verification using deep cnn features*,CoRR, abs/1508.01722, 2015.

[26] S. Sankaranarayanan, A. Alavi, and R. Chellappa, *Triplet similarity embedding for face verification*, arxiv preprint, arXiv:1602.03418, 2016

[27] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman, *From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose*,In: PAMI, 2001

[28] A.M. Martinez and R. Benavente, *The AR Face Database*, CVC Technical Report #24, June 1998

[29] *Evaluation of Clustering*, `https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html`

[30] Lior Wolf, Tal Hassner and Itay Maoz, *Face Recognition in Unconstrained Videos with Matched Background Similarity*, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2011

[31] *A. Vedaldi and B. Fulkerson. VLFeat library*, http://www.vlfeat.org/,2008.

[32] D. G. Lowe., *Distinctive image features from scale-invariant keypoints*, IJCV, 60(2):91110, 2004.

[33] N. Dalal, B. Triggs, *Histograms of oriented gradients for human detection*, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1-8, Jun. 2005.

[34] Ahonen, T., Hadid, A., and Pietikinen, M., *Face description with local binary patterns: Application to face recognition*, In: PAMI 2006

[35] Philipp Wagner , *Geometric Warping* http://www.bytefish.de/blog/aligning_face_images/

[36] Christopher J. C. Burges, *A tutorial on support vector machines for pattern recognition*, In: Data Mining and Knowledge Discovery, 1997

[37] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma,*Robust Face Recognition via Sparse Representation* , In: PAMI 2009

[38] M. Turk and A. Pentland, *Eigenfaces for Recognition*, J. Cognitive Neuroscience, vol. 3, no. 1, 1991.

[39] Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. 1997, *Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection*, IEEE Trans. Patt. Anal. Mach. Intell. 19, 711720.