Domain-Specific Face Synthesis for Still-to-Video Face Recognition

by

Faniya MOKHAYYERI

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, JANUARY 24, 2020

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Prof. Eric Granger, Thesis Supervisor
Department of Systems Engineering, École de Technologie Supérieure

Prof. Luc Duong, President of the Board of Examiners
Department of Systems Engineering, École de Technologie Supérieure

Prof. Christian Desrosiers, Member of the jury
Department of Systems Engineering, École de Technologie Supérieure

Prof. Nizar Bouguila, External Examiner
Concordia Institute for Information Systems Engineering, Concordia University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON "JANUARY 24, 2020"

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to express my appreciation and thanks to my supervisor, Prof. Eric Granger, whose guidance added considerably to my graduate experience. His invaluable advice and careful editing contributed enormously to the production of this thesis, and without his support, it would not be possible to conduct this research.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Christian Desrosiers, Prof. Luc Duong and Prof. Nizar Bouguila.

Many thanks to my fellow labmates at the LIVIA for making the effort to stay in touch and all the good times that were a necessary break from my normal Ph.D. life.

I wish to express my deep gratitude to my parents for their unconditional love and lifelong support, and for making me the way I am today. I would also like to thank my younger brother who in his own way gave me encouragement and joy to move on.

My sincere thanks goes to my beloved husband, Kaveh, for his unbelievable support and encouragement every step of the way. I truly thank him for sticking by my side during my good and bad times. This thesis is dedicated to him.

GÉNÉRATION DE VISAGES SYNTHÉTIQUES POUR LA RECONNAISSANCE DE VISAGES SUR VIDÉO

Faniya MOKHAYYERI

RÉSUMÉ

La reconnaissance faciale (FR) en vidéo continue de susciter un intérêt considérable de la part des milieux universitaires et de l'industrie en raison du large éventail d'applications dans les domaines de la surveillance et de la sécurité. Malgré les progrès récents en matière de vision par ordinateur et d'apprentissage machine, la conception d'un système robuste pour de reconnaissance faciale en temps réel pour les applications de surveillance reste un défi important. Un problème clé est la divergence entre les visages du domaine source, où les visages de référence sont de haute qualité et capturés dans des conditions contrôlées par des caméras fixes, et ceux du domaine cible, où les images vidéo sont capturées avec des caméras vidéo dans des conditions non contrôlées avec des variations de pose, éclairage, flou, etc. L'apparence des visages capturés dans les vidéos correspond à de multiples distributions de données pouvant différer considérablement des visages initialement capturés. Un autre défi de la vidéo est le nombre limité de photos de référence disponibles par personne cible pour la conception de modèles de visage. Ce scénario est courant dans les applications de sécurité et de surveillance basées sur la vidéo, comme par exemple l'authentification biométrique et le triage avec une liste de surveillance. Les performances des systèmes vidéos peuvent diminuer considérablement en raison de la quantité limitée d'information disponible pour représenter les variations intra-classe observées dans les images.

Cette thèse propose des techniques d'augmentation des données basées sur la synthèse des visages pour surmonter les défis posés par la variation des visages et le nombre limité d'images d'entraînement. Le principal avantage des approches proposées est la possibilité de fournir un ensemble compact capable de représenter avec précision le visage de référence d'origine avec des variations pertinentes aux condition de capture dans le domaine cible. En particulier, cette thèse présente 3 nouveaux systèmes pour une reconnaissance faciale robuste en vidéo qui sont basés sur l'augmentation synthétique des galeries de référence.

Dans une première contribution, une approche de synthèse de visage exploitant les informations de variation représentatives intra-classe du domaine cible est proposée. Cette approche, appelée synthèse de visages spécifique à un domaine, génère un ensemble compact de visages synthétiques qui ressemblent à des individus d'intérêt dans les conditions de capture pertinentes pour le domaine cible. Dans une implémentation particulière basée représentation clairsemée, les visages synthétiques générés sont utilisés pour former un dictionnaire interdomaine tenant compte de la structure de la clarté, où les blocs de dictionnaire combinent les visages d'origine et synthétique de chaque individu. Les résultats expérimentaux obtenus avec des vidéos des bases de données Chokepoint et COX-S2V révèlent qu'augmenter le nombre de galeries de référence de systèmes la FR en vidéo en utilisant l'approche proposée par une approche synthèse de visage peut fournir un niveau de précision nettement supérieur à celui de l'état de l'art.

Dans un deuxième temps, nous proposons un modèle de représentation par paires fragmentées permettant l'utilisation d'informations conjointe variationnelles et d'images de visage synthétiques. Le modèle proposé, appelé *modèle de synthèse plus variationnel*, reconstruit une image sonde en utilisant conjointement (1) un dictionnaire variationnel conçu avec un ensemble générique et (2) un dictionnaire de galerie complété par un ensemble d'images synthétique générées sur une grande diversité des angles de pose. Le dictionnaire de galerie augmentée est ensuite encouragé à partager le même motif de parcimonie avec le dictionnaire de variation pour d'angles de pose similaires en résolvant un problème d'optimisation simultané basé sur la parcimonie. Les résultats expérimentaux obtenus sur les données Chokepoint et COX-S2V, indiquent que l'approche proposée peut surpasser les méthodes représentation clairsemé de pointe pour la FR en vidéo continue avec un seul échantillon par personne.

Troisièmement, un réseau siamois profond, appelé SiamSRC, est proposé pour effectuer une mise en correspondance des visages à l'aide d'une représentation clairsemée. L'approche proposée étend la galerie en utilisant un ensemble d'images de visage synthétiques et exploite la représentation clairsemé avec une structure de blocs pour la correspondance des visages par paires qui trouve la représentation d'une image sonde nécessitant le nombre minimal de blocs de la galerie. Les résultats expérimentaux obtenus avec les bases de données Chokepoint et COX-S2V suggèrent que le réseau SiamSRC proposé permet une représentation efficace des variations intra-classe avec une augmentation modérée de la complexité temporelle. Les résultats ont montré que les performances des systèmes d'images fixes à vidéo continue basées sur SiamSRC peuvent être améliorées grâce à la synthèse des visages, sans qu'il soit nécessaire de collecter une grande quantité de données d'entraînement.

Des expérimentations approfondies ont été menées sur deux ensembles de données de surveillance disponibles au public. Les résultats ont indiqué que la synthèse de visage à elle seule ne peut pas résoudre efficacement les défis du échantillons limités et les problèmes de changement de domaine visuel. Les techniques proposées, à savoir l'intégration de la synthèse des visages et de l'apprentissage générique, peuvent fournir un niveau de précision supérieur à celui des approches de pointe avec un seul échantillon par personne.

Mots-clés: Reconnaissance de visage, synthèse de visage, reconstruction du visage en 3D, surveillance vidéo, adaptation de domaine, représentation clairsemée, apprentissage générique

DOMAIN-SPECIFIC FACE SYNTHESIS FOR STILL-TO-VIDEO FACE RECOGNITION

Faniya MOKHAYYERI

ABSTRACT

Face recognition (FR) has attracted a considerable amount of interest from both academia and industry due to the wide range of applications as found in surveillance and security. Despite the recent progress in computer vision and machine learning, designing a robust system for video-based FR in real-world surveillance applications has been a long-standing challenge. One key issue is the visual domain shift between faces from source domain, where high-quality reference faces are captured under controlled conditions from still cameras, and those from the target domain, where video frames are captured with video cameras under uncontrolled conditions with variations in pose, illumination, expression, etc. The appearance of the faces captured in the videos corresponds to multiple non-stationary data distributions can differ considerably from faces captured during enrollment. Another challenge in video-based FR is the limited number of reference stills that are available per target individual to design facial models. This is a common scenario in security and surveillance applications, as found in, e.g., biometric authentication and watch-list screening. The performance of video-based FR systems can decline significantly due to the limited information available to represent the intra-class variations seen in video frames. This thesis proposes 3 data augmentation techniques based on face synthesis to overcome the challenges of such visual domain shift and limited training set. The main advantage of the proposed approaches is the ability to provide a compact set that can accurately represent the original reference face with relevant intra-class variations corresponding to the capture conditions in the target domain. In particular, this thesis presents new systems for domain-invariant still-to-video FR that are based on augmenting the reference gallery set synthetically which are described with more details in the following.

As a first contribution, a face synthesis approach is proposed that exploits the representative intra-class variational information available from the generic set in target domain. The proposed approach, called domain-specific face synthesis, generates a set of synthetic faces that resemble individuals of interest under the capture conditions relevant to the target domain. In a particular implementation based on sparse representation, the generated synthetic faces are employed to form a cross-domain dictionary that accounts for structured sparsity where the dictionary blocks combine the original and synthetic faces of each individual. Experimental results obtained with videos from the Chokepoint and COX-S2V datasets reveal that augmenting the reference gallery set of still-to-video FR systems using the proposed face synthesizing approach can provide a significantly higher level of accuracy compared to state-of-the-art approaches.

As a second contribution, a paired sparse representation model is proposed allowing for joint use of generic variational information and synthetic face images. The proposed model, called synthetic plus variational model, reconstructs a probe image by jointly using (1) a variational dictionary designed with generic set and (2) a gallery dictionary augmented with a set of synthetic images generated over a wide diversity of pose angles. The augmented gallery dictionary is then encouraged to share the same sparsity pattern with the variational dictionary for similar pose angles by solving a simultaneous sparsity-based optimization problem. Experimental results obtained on Chokepoint and COX-S2V datasets, indicate that the proposed approach can outperform state-of-the-art methods for still-to-video FR with a single sample per person.

As a third contribution, a deep Siamese network, referred as SiamSRC, is proposed where performs face matching using sparse coding. The proposed approach extends the gallery using a set of synthetic face images and exploits sparse representation with a block structure for pairwise face matching that finds the representation of a probe image that requires the minimum number of blocks from the gallery. Experimental results obtained using the Chokepoint and COX-S2V datasets suggest that the proposed SiamSRC network allows for efficient representation of intra-class variations with only a moderate increase in time complexity. Results show that the performance of still-to-video FR systems based on SiamSRC can improve through face synthesis, with no need to collect a large amount of training data.

Results indicate that our proposed techniques which are the integration of face synthesis and generic learning can effectively resolve the challenges of the visual domain shift and limited number of reference stills and provide a higher level of accuracy compared to state-of-the-art approaches under unconstrained surveillance conditions.

Keywords: Face Recognition, Face Synthesis, 3D Face Reconstruction, Video Surveillance, Domain Adaptation, Sparse Representation, Generic Learning

TABLE OF CONTENTS

Page

INTRODUCTION 1

CHAPTER 1 LITERATURE REVIEW 11
1.1 Still-to-Video Face Recognition Systems 11
1.1.1 Challenges 12
1.2 Data Augmentation 13
1.2.1 Face Synthesis 15
1.2.1.1 3D Morphable Model 15
1.2.1.2 Generative Adversarial Network 17
1.2.2 Generic Learning 18
1.3 Deep Face Recognition 19
1.4 Summary 21

CHAPTER 2 DOMAIN-SPECIFIC FACE SYNTHESIS FOR VIDEO FACE RECOGNITION FROM A SINGLE SAMPLE PER PERSON 23
2.1 Introduction 24
2.2 Related Work 27
2.2.1 Multiple Face Representations 28
2.2.2 Generic Learning 28
2.2.3 Synthetic Face Generation 29
2.3 Domain-Specific Face Synthesis 31
2.3.1 Characterizing the Capture Conditions 31
2.3.1.1 Estimation of Head Pose 33
2.3.1.2 Luminance-Contrast Distortion 33
2.3.1.3 Representative Selection 34
2.3.2 Face Synthesis 37
2.3.2.1 Intrinsic Image Decomposition 37
2.3.2.2 3D Face Reconstruction 38
2.3.2.3 Illumination Transferring 39
2.4 Domain-invariant Face Recognition with DSFS 40
2.5 Experimental Methodology 44
2.5.1 Databases 44
2.5.2 Experimental protocol 45
2.5.3 Performance Measures 46
2.6 Results and Discussion 47
2.6.1 Face Synthesis 47
2.6.1.1 Frontal View 50
2.6.1.2 Profile View 50
2.6.2 Face Recognition 51
2.6.2.1 Pose Variations 51

2.6.2.2 Mixed Pose and Illumination Variations 51
2.6.2.3 Impact of Representative Selection 53
2.6.3 Comparison with Reference Techniques 57
2.6.3.1 Generic Set Dimension 57
2.7 Conclusions 60

CHAPTER 3 A PAIRED SPARSE REPRESENTATION MODEL FOR ROBUST FACE RECOGNITION FROM A SINGLE SAMPLE 63
3.1 Introduction 64
3.2 Background on Sparse Coding 68
3.2.1 Sparse Representation-based Classification 69
3.2.2 SRC through Generic Learning 69
3.3 The Proposed Approach - A Synthetic plus Variational Model 72
3.3.1 Dictionary Design 73
3.3.2 Synthetic Plus Variational Encoding 76
3.4 Face Recognition with the S+V Model 79
3.5 Experimental Methodology 81
3.5.1 Datasets 81
3.5.2 Protocol and Performance Measures 83
3.6 Results and Discussion 84
3.6.1 Synthetic Face Generation 84
3.6.2 Impact of Number of Synthetic Images 86
3.6.3 Impact of Camera Viewpoint 88
3.6.4 Impact of Feature Representations 89
3.6.5 Comparison with State-of-the-Art Methods 91
3.6.6 Ablation Study 95
3.6.7 Complexity Analysis 96
3.7 Conclusion 97

CHAPTER 4 VIDEO FACE RECOGNITION USING SIAMESE NETWORKS WITH BLOCK-SPARSITY MATCHING 99
4.1 Introduction 100
4.2 Related Work 103
4.2.1 Deep Siamese Networks for Face Recognition 103
4.2.2 Face Synthesis 105
4.2.3 Sparse Representation-based Classification 106
4.3 The SiamSRC Network 107
4.3.1 Notation 108
4.3.2 Representative Selection 108
4.3.3 Face Synthesis 110
4.3.4 Block-Sparsity Matching 111
4.4 Experimental Methodology 115
4.4.1 Datasets 115
4.4.2 Protocol and Performance Measures 116

4.5 Results and Discussion 118
4.5.1 Synthetic Face Generation 118
4.5.2 Impact of Number of Synthetic Images 119
4.5.3 Comparison with State-of-the-Art Methods 123
4.5.4 Ablation Study 124
4.6 Conclusion 126

CONCLUSION AND RECOMMENDATIONS 127

LIST OF PUBLICATIONS 131

APPENDIX I CROSS-DOMAIN FACE SYNTHESIS USING A CONTROLLABLE GENERATIVE ADVERSARIAL NETWORK 133

BIBLIOGRAPHY 152

LIST OF TABLES

Page

Table 2.1 Average DSQ value for frontal and profile views on Chokepoint and COX-S2V datasets. 51

Table 2.2 Average error rate of pose estimation for frontal and profile views on Chokepoint and COX-S2V datasets. 56

Table 2.3 Comparative transaction level analysis of the proposed FR approach and related state-of-the art FR methods with Chokepoint and COX-S2V databases. 58

Table 2.4 Average computational complexity of the DSFS and state-of-the-art methods on Chokepoint and COX datasets. 60

Table 3.1 Average accuracy of FR systems based on the proposed S+V model, SRC, and ESRC over different sessions, portals and cameras of the Chokepoint dataset. Feature representations are raw pixels, the 3DMM method is used for face synthesis. 90

Table 3.2 Average accuracy of FR systems using the proposed S+V model, SRC, and ESRC over different sessions and portals of the COX-S2V dataset. Feature representations are raw pixels, the 3DMM method is used for face synthesis. 90

Table 3.3 Average accuracy of FR systems using the proposed S+V model and template matching using different feature representation on Chokepoint and COX-S2V databases. 91

Table 3.4 Average accuracy of FR systems using the proposed S+V model (3DMM face synthesis) with different deep feature representations on Chokepoint and COX-S2V databases. 92

Table 3.5 Average accuracy of FR systems using the augmented dictionary (3DMM face synthesis) and auxiliary dictionaries on Chokepoint and COX-S2V databases. 92

Table 3.6 Average accuracy of FR systems based on the proposed S+V model and related state-of-the art SRC methods for videos from all 3 cameras of the Chokepoint and COX-S2V databases. Feature representations are raw pixels, the 3DMM method is used for face synthesis. 93

Table 3.7 Results of ablation study with Chokepoint database. 96

Table 3.8 Results of ablation study with COX-S2V database. 96

Table 3.9 Average time required by techniques to classify a probe videos ROI with the Chokepoint and COX-S2V datasets. 97

Table 3.10 Average computational time of different step in the S+V model with the Chokepoint and COX-S2V datasets. 97

Table 4.1 Average accuracy of the SiamSRC network and related state-of-the networks for videos from all 3 cameras of the Chokepoint and COX-S2V databases. Feature representations are raw pixels, the 3DMM method is used for face synthesis. 125

Table 4.2 Average matching time of deep Siamese networks over the video ROIs of the Chokepoint and COX-S2V datasets. 125

Table 4.3 Ablation study results with Chokepoint dataset. 126

Table 4.4 Ablation study results with COX-S2V dataset. 126

LIST OF FIGURES

Page

Figure 0.1 The flow of the thesis is shown by connected boxes. The solid arrows indicate the dependencies between the chapters and appendix (i.e., one chapter must be read beforehand). Dashed arrows indicates the suggested readings between the chapters and appendix for a better comprehension. 8

Figure 1.1 Taxonomy of the solutions proposed in the literature to address challenging issues of FR systems. 14

Figure 2.1 Overview of the proposed DSFS algorithm to augment the reference gallery set. We assume that the gallery set initially contains only one reference still image per individual of interest. 25

Figure 2.2 Block diagram of the DSFS technique applied to a reference still ROI. 32

Figure 2.3 Pipeline for characterizing capture conditions of video ROIs in the target domain. 33

Figure 2.4 An illustration of the AP clustering process. 35

Figure 2.5 Block diagram of the proposed domain-invariant SRC-based still-to-video FR system. 41

Figure 2.6 Examples of representative selection results using AP clustering technique in terms of pose angles with Chokepoint dataset on video sequences of 10 individuals and COX-S2V dataset on video sequences of 100 individuals, respectively. 48

Figure 2.7 Examples of luminance and contrast representatives with Chokepoint dataset on video sequences of 10 individuals and COX-S2V dataset on video sequences of 100 individuals, respectively, where the center of clusters show exemplars. 48

Figure 2.8 Examples of synthetic ROIs generated under different capture conditions using the DSFS technique with Chokepoint (a,b) and COX-S2V (c,d) datasets. 49

Figure 2.9 Synthetic face images generated under different pose via (a) DSFS, (b) 3DMM, (c) 3DMM-CNN with Chockpoint dataset. 50

Figure 2.10 Average AUC and AUPR versus the number of synthetic ROIs generated with DSFS according to various pose and fixed illumination. The still-to-video FR system employs either SVM and SRC classifiers on Chokepoint (a,b) and COX-S2V (c,d) databases. 52

Figure 2.11 Average AUC and AUPR versus the number of synthetic ROIs generated with DSFS, 3DMM, and SHBMM according to pose and lighting effects where still-to-video FR system employs either SVM and SRC classifiers on Chokepoint (a,b) and COX-S2V (c,d) databases. 53

Figure 2.12 Average AUC and AUPR of a still-to-video FR system designed with representative synthetic ROIs vs a system designed with randomly generated synthetic ROIs on Chokepoint (a,b) and COX-S2V (c,d) datasets. 55

Figure 2.13 Average AUC and AUPR accuracy obtained by increasing the number of the synthetic ROIs generated using DSFS from different representative pose angles (obtained with different pose estimation techniques) on Chokepoint (a,b) and COX-S2V (c,d) datasets. 56

Figure 2.14 Average AUC and AUPR accuracy obtained by increasing the size of the generic set in the (synthetic) variant dictionary on Chokepoint (a,b) and COX-S2V (c,d) datasets. 59

Figure 2.15 Time complexity versus the number of synthetic ROIs on Chokepoint (a) and COX-S2V (b) data. 59

Figure 3.1 Overall architecture of the proposed approach. The gallery dictionary is augmented with a diverse set of synthetic images and the auxiliary variational dictionary co-jointly encode non-linear variations in appearance. Sparse coefficients within each dictionary share the same sparsity pattern in terms of pose angle. 66

Figure 3.2 A comparison of the coefficient matrices for three sparsity models: (a) Independent sparsity (SRC) with a single dictionary, (b) Extended sparsity (ESRC) with two dictionaries, and (c) Paired extended sparsity (S+V model) with pair-wise correlation between two dictionaries where the sparse coefficients of same poses share the same sparsity pattern. Each column represents a sparse coefficient vector and each square block denotes a coefficient value. White blocks denote zero values, whereas color blocks stand for nonzero values. 73

Figure 3.3 An illustration of sparsity pattern with the S+V model based on clustering results in the pose space. Each column represents a sparse representation vector, each square denotes a coefficient and each matrix is a dictionary. 77

Figure 3.4 Block diagram of the proposed still-to-video FR system with the S+V modeling. 80

Figure 3.5 Examples of still images and video frames from portals and cameras of Chokepoint dataset. 82

Figure 3.6 Examples of still images and video frames from 3 cameras of COX-S2V dataset. 82

Figure 3.7 Example of clusters obtained with 20 and 40 facial trajectories represented in the pose space with Chokepoint (ID#1, #5, #6, #7, #9) and COX-S2V (ID#1, #9, #24, #33, #36, #38, #44, #56, #78, #80) datasets, respectively. Clusters are shown with different colors, and their representative pose exemplars are indicated with a black circle. 85

Figure 3.8 Examples of synthetic face images generated from the reference still ROI of individuals ID#25 and ID#26 (a) of Chokepoint dataset. They are produced based on representative exemplars (poses) and using 3DMM (b) and 3DMM-CNN (c). 86

Figure 3.9 Examples of synthetic face images generated from the reference still ROI of individuals ID#21 and ID#151 (a) of COX-S2V dataset. They are produced based on representative exemplars (poses) and using 3DMM (b) and 3DMM-CNN (c). 87

Figure 3.10 Average pAUC(20%) and AUPR accuracy of S+V model versus the size of the synthetic set generated using 3DMM and 3DMM+CNN on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation. 88

Figure 3.11 Average pAUC(20%) and AUPR accuracy for SRC, ESRC and S+V model on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation. 89

Figure 3.12 Illustration of procedure for the selection of the largest pose variations. 93

Figure 3.13 Average pAUC(20%) and AUPR accuracy of S+V model and related state-of-the-art techniques versus the different pose variations on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation. 94

Figure 3.14 Average pAUC(20%) and AUPR accuracy versus the size of the generic set on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation. 95

Figure 4.1 An overview of the proposed deep SiamSRC network for still-to-video FR. 102

Figure 4.2 Architecture of SiamSRC. 113

Figure 4.3 The block diagram that summarizes the steps for design and inference with the proposed SiamSRC network. 115

Figure 4.4 Example of clusters obtained with 20 and 40 facial trajectories represented in the pose space with Chokepoint (a) and COX-S2V (b) datasets, respectively. Clusters are shown with different colors, and their representative exemplars are indicated with a black circle. 119

Figure 4.5 Examples of the reference still ROI of individuals ID#25 and ID#26 of Chokepoint dataset (a), ID#21 and ID#151 of COX-S2V dataset (c), and their corresponding synthetic face images (b,d) produced using the representatives and domain-specific 3DMM. 119

Figure 4.6 Average pAUC(20%) and AUPR accuracy of SiamSRC network versus the size of the synthetic set on Chokepoint (a,b) and COX-S2V (c,d) databases. Synthetic faces are generated using 3DMM, domain-specific 3DMM, and 3DMM-CNN. Error bars are standard deviation. 120

Figure 4.7 Average pAUC(20%) and AUPR accuracy of SiamSRC network versus the size of the synthetic set on Chokepoint (a,b) and COX-S2V (c,d) databases. Synthetic faces are generated using 3DMM, domain-specific 3DMM, and 3DMM-CNN. Error bars are standard deviation. 121

Figure 4.8 Average pAUC(20%) and AUPR accuracy for SRC model, and Siamese and SiamSRC networks with q representative and 15 randomly selected faces on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation. 122

Figure 4.9 Average pAUC(20%) and AUPR accuracy versus the size of the generic set on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation. 123

Figure 4.10 Average pAUC(20%) and AUPR accuracy for S+V model and SiamSRC network of using an augmented gallery with both generic set and synthetic faces on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation. 124

LIST OF ALGORITHMS

Page

Algorithm 2.1 The DSFS Approach. 40

Algorithm 2.2 A SRC-based Still-to-Video FR System. 44

Algorithm 3.1 Synthetic Plus Variational Model. 79

Algorithm 4.1 SiamSRC Network. 114

LIST OF ABREVIATIONS

ADM	Alternating Direction Method
AUC	Area Under Curve
AP	Affinity Propagation
AUPR	Area Under Precision-Recall
CNN	Convolutional Neural Network
CSR	Customized Sparse Representation-based Classification
DA	Domain Adaptation
DSFS	Domain Specific Face Synthesis
DSQ	Domain Shift Quantification
ESRC	Extended Sparse Representation-based Classification
FR	Face Recognition
FPR	False Positive Rate
GAN	Generative Adversarial Model
GLQ	Global Luminance Distortion in Image Quality
GCQ	Global Contrast Distortion in Image Quality
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
3DMM	3D Morphable Model
3DMM-CNN	CNN Regresssed 3D Morphable Model

PCA	Principal Component Analysis
pAUC	Partial Area Under Curve
RADL	Robust Auxiliary Dictionary Learning
ROC	Receiver Operating Characteristic
ROI	Region of Interest
LGR	Local Generic Representation
S+V	Synthetic plus Variational Model
SCI	Sparsity Concentration Index
SHBMM	Spherical Harmonic Basis Morphable Model
SRC	Sparse Representation-based Classification
SSPP	Single Sample Per Person
SVDL	Sparse Variation Dictionary Learning
SVM	Support Vector Machine
TM	Template Matching
TP	True Positive Rate

LISTE OF SYMBOLS AND UNITS OF MEASUREMENTS

r	Reference still ROI
y	Probe ROI
g	Generic ROI
s	Synthetic ROI
p	Head pose angle
R	Reference set
S	Augmented set
D	Gallery dictionary
V	Variational dictionary
D'	Augmented gallery dictionary
l	Global luminance distortion
c	Global contrast distortion
s_u	Illumination-contrast similarities
C_c	Positive stabilizing constant
σ_k	Standard deviation
$\mathbf{U}_{ji}$	Illumination-contrast cluster
θ^{pitch}	Pitch rotation
θ^{yaw}	Yaw rotation
θ^{roll}	Roll rotation

$\hat{s}_k$	Basis vector
λ	Sparsity regularization parameter
α	Sparse vector of reference coefficient
β	Sparse vector of generic coefficients
A	Joint Sparse vector of reference coefficient
B	Joint Sparse vector of generic coefficients
$\mathit{\Lambda}$	Orthographic projection matrix
η	Rank regularization parameter
ξ	Sparsity level
τ	Rejection threshold
n_d	Number of patches
d_p	Feature dimension of patches
K	Number of clusters
ψ	Sliding step
L	Dynamic range of the pixel values
r_p	Pose responsibility
q	Number of synthetic samples
W	Weigh of clusters
h_d	Identity label
t_{2d}	Translation vector

R_{enc}	Encoder
R_{dec}	Decoder
z	Noise model
D_R	Discriminator
D_F	Feature-based discriminator
C	Classifier
h	Simulated set labels
θ_R	Refiner parameters
θ_F	Feature extractor parameters
θ_F	Classifier parameters
θ_F	Classifier parameters
$\mathcal{L}_D$	Adversarial loss for refiner
$\mathcal{L}_{D_F}$	Adversarial loss for feature extractor
$\mathcal{L}_C$	Classifier loss
$\mathcal{L}_F$	Feature loss
$\mathcal{L}_{real}$	Realism loss
$\mathcal{L}_{reg}$	Identity preserving loss

INTRODUCTION

Given the growing demand for public security, much attention has been devoted to person identification and recognition. Over the past few decades, biometric technology has received a lot of interest as a way for public security agencies to achieve accurate identification and verification. Among biometric applications, Face Recognition (FR) in video surveillance applications is considered as a promising approach for authentication owing to its convenient data collection, flexible control, high performance to cost ratio as well as the possibility of analysis of live feeds. FR has been the prominent biometric modality for identity authentication and has been widely used in many areas, such as military, finance, public security and daily life. It is reported that the global market for video surveillance technologies has reached revenues in the billions of US$ as traditional analogue technologies are replaced by IP-based digital ones (Pagano *et al.* (2014)). Hence, the necessity to develop an efficient and robust FR in surveillance videos is rising. Using video streams makes it possible to employ spatial and temporal information of faces to improve FR performance. Furthermore, some on-line and incremental learning techniques can be applied for video-based FR to update the model over time. The growing availability of low-cost cameras also motivated the development of intelligent video surveillance system based on FR algorithms (Huang *et al.* (2017b); Cevikalp *et al.* (2019)).

FR from video surveillance can be designed using two scenarios w.r.t. the nature of training and testing data: (1) video-to-video, (2) still-to-video. In V2V scenario, training and recognition tasks are performed using frames from video streams. However, in still-to-video scenario, still images of individuals are used to design facial models, and recognition is performed using frames from video streams (Dewan *et al.* (2016)). A generic still-to-video FR consists of two main phases; enrollment and operation. During the enrollment of target individuals, facial models are designed using facial regions of interests (ROIs) isolated in high quality reference still images that were captured under controlled conditions. During operations, the ROIs of faces captured with surveillance cameras under uncontrolled conditions are compared against the facial models of watch-list individuals. A face tracker may be employed to track the ROIs appeared in the capturing scene over several frames, and matching scores can be accumulated over a facial trajectory for robust spatio-temporal recognition (Bashbaghi *et al.* (2017a)). The typical application of still-to-video FR is watch-list screening where ROIs captured over a network of video surveillance cameras are matched against ROIs extracted from reference stills of target individuals (high quality mug shots) that were captured under controlled condition (De-la Torre *et al.* (2015)). Recently, the theory of sparse representation and compressed sensing has shown promising results on this challenging problem. The basic idea is to cast the recognition problem as one of classifying among multiple linear regression models. Given sufficient training samples from each class, it will be possible to represent the test samples as a linear combination of those training samples from the same class. These algorithms produce extremely striking results and accurately recognize subjects across large databases despite severe corruption and occlusion (Gao *et al.* (2017)). Using deep neural networks to learn effective feature representations has become popular in FR Deng *et al.* (2019); Masi *et al.* (2019a); Wu *et al.* (2018b) and FR accuracy has been boosted rapidly in recent years. Thanks to their deep architectures and large learning capacity, effective features can be learned through hierarchical nonlinear mappings. Contemporary deep models report near perfect performance on challenging benchmarks such as Labeled Faces in the Wild (Huang *et al.* (2008)), which due to its difficulty, represented the de facto standard for evaluating FR technology for nearly a decade (Jiang *et al.* (2019)).

Problem Statement

This thesis focuses on video-based FR, in particular still-to-video FR, where a single high-quality reference still image under controlled conditions is matched against lower-quality faces captured with video cameras under uncontrolled conditions.

Numerous performance evaluation have shown that FR algorithms that operate well in controlled environments tend to suffer in unconstrained conditions (Sohn *et al.* (2017)). For a still-to-video FR system, critical obstacles towards surveillance applications are often caused by large intra-class variability, arising from changes in lighting, pose, expression and corruption. Another key issue is the limited number of reference stills that are available per target individual to design facial models. In many surveillance applications (e.g., watch-list screening), only a single reference still per person is available for design, which corresponds to the so-called Single Sample Per Person (SSPP) problem. A further challenge for surveillance applications is the matching of low-resolution probe face images with high resolution reference images, which could be the case in watch-list scenarios (Li *et al.* (2018)). Over the past few decades, a wide variety of approaches have emerged to overcome the aforementioned problems that affect FR performance such as; (1) multiple face representations which extract discriminant features from face images that are robust to facial variations (Bashbaghi *et al.* (2017b); Yin *et al.* (2019)), (2) image patching which decomposes an image into multiple local components to provide robustness to local changes in illumination, blur, etc (Zhang *et al.* (2018b)), (3) super resolution that attempts to obtain a high-resolution face image by leveraging the knowledge of multiple low-resolution images (Yu *et al.* (2018)), (4) frontalization[1] and illumination normalization that try to adjust the images to normal pose and lighting condition (Cao *et al.* (2019b)), (5) generic learning (incorporating generic auxiliary set) (Deng *et al.* (2012)) where a genetic training set[2] is used to exploit variational information from an auxiliary generic set of images to represent the differences between probe and reference gallery images (Wei & Wang (2015); Deng *et al.* (2018)), (6) face synthesis which attempt to generate synthetic face images from the original reference stills under different appearance (such as pose, illumination, expression, and etc.) (Sanyal *et al.* (2019); Masi *et al.* (2016)) and add the extra samples to the gallery to

[1] Frontalization is the process of synthesizing frontal facing views of faces appearing in single unconstrained photos.

[2] A generic set is defined as an auxiliary set comprised of facial ROIs from unknown individuals captured in videos from the target environment.

produce diverse face representations and accordingly improve the robustness to various capture conditions.

This Thesis focuses on methods that are based on augmenting the reference gallery set through synthetic set, and by taking into account the intra-class variation information transferred from a generic set. The most general approach for the generation of synthetic face images under different appearance is 3D reconstruction that attempts to reconstruct 3D models of faces from their 2D images. 3D morphable model (3DMM) and its variants have been the most popular methods for 3D face reconstruction for many years. However, the resulting images are not realistic enough to be suitable for real-world FR tasks. The discrepancy in quality of synthetic and real images has been the main challenge of using synthetic data obtained from 3DMM model (Gecer *et al.* (2019), Tran & Liu (2019), Koppen *et al.* (2018)). Recently, generative adversarial networks (GANs) have been successful to mitigate such challenges. Zhao & et al. (2017) proposed Dual-GAN which improves the realism of a 3DMM's output using unlabeled real faces, while preserving the identity information during the refinement. The dual agents are designed for distinguishing real v.s. fake and identities simultaneously. Gecer *et al.* (2018) proposed an end-to-end semi-supervised adversarial framework to synthesize photorealistic images conditioned by synthetic images generated by 3DMM with a wide range of expressions, poses, and illuminations. Despite their success, in practice, GANs are likely to get stuck in mode collapse for large scale image generation.

Selecting a sufficient number of faces to cover intra-class variations is another challenging issue of data augmentation techniques. Many synthetic faces may be generated to account for all possible capture conditions. In this case, FR systems would, therefore, require complex implementations and may yield lower accuracy when training on many facial images that provide less relevant information for FR.

Research Objectives and Contributions

The main objective of this thesis is to develop an accurate and robust still-to-video FR system that can accurately recognize target individuals under real-world capture conditions when only one reference face image per person is available for facial modeling. According to the constraints of real-world surveillance applications, these systems need to be designed considering only a single reference still captured under controlled conditions, while they should be operated over the low-quality videos captured under uncontrolled conditions such as various pose, lighting, expression, and etc. We first study what are the representative information to compensate the real-world intra-class variations. Then, we propose domain specific face synthesis methods that take into account the representative information obtained from a generic set during the synthesis process. The generated synthetic images are used to enrich the reference gallery. Our problem of generating domain-specific face images can be seen as a domain adaptation[3] problem *i.e.* aligning reference stills of source domain into probe videos of target domain (Hong *et al.* (2017)).

Since this thesis is article-based, each chapter presents a different contribution to the development of the robust still-to-video FR framework. There are three main contributions in this work which led to three journal and three conference publications. The contributions of this thesis are listed below:

I) An algorithm for domain-specific face synthesis is proposed. It maps representative variation information from the generic set in the target domain to the reference stills by integrating an image-based face relighting technique inside the 3D reconstruction framework. In order to find a representative set of faces, affinity propagation clustering is applied in

[3] Domain adaptation tries to learn a better model for the target scenario, by on one hand borrowing some common knowledge from the source domain and on the other hand exploiting the particular information from the target domain but with limited supervision

the captured condition space defined by pose and illumination estimation. In this way, a compact set of synthetic faces is produced that represent reference images and probe video frames under a common capture condition.

Related publications:

- Mokhayeri, F., Granger, E., and Bilodeau, G-A. "Domain-specific face synthesis for video face recognition from a single sample per person." IEEE Transactions on Information Forensics and Security 14.3 (2018): 757-772.
- Mokhayeri, F., Granger, E., and Bilodeau, G-A. "Synthetic face generation under various operational conditions in video surveillance." In IEEE International Conference on Image Processing (ICIP), 2015.

II) A paired sparse representation framework for FR is introduced that reconstructs a probe image using an augmented gallery dictionary enriched with a set of synthetic stills generated under a wide diversity of pose angles and auxiliary dictionary designed with generic set. Two dictionaries are correlated by imposing the simultaneous sparsity prior that force the augmented dictionary to pair the same sparsity pattern with auxiliary dictionary for the same pose angles.

Related publications:

- Mokhayeri, F., Granger, E. "A paired sparse representation model for robust face recognition from a single sample." Pattern Recognition 100 (2020): 107-129.
- Mokhayeri, F., Granger, E. "Robust video face recognition from a single still using a synthetic plus variational model." In IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2019.

III) Deep SiamSRC network is proposed that employs sparse representation with block structure for face matching inside a Siamese network. In this approach, a set of domain-specific synthetic facial images are generated and then integrated into the reference gallery of the Siamese network where rendering parameters are obtained through row sparsity clustering of unlabeled faces.

Related publications:

- Mokhayeri, F., Granger, E. "Video face recognition using siamese networks with block-sparsity matching." IEEE Transactions on Biometrics, Behavior, and Identity Science (2019).

Organization of the Thesis

This is a thesis by articles, with Chapters 2 to 4 corresponding to a journal paper. Figure 0.1 presents an overview of the organization of this thesis. The chapters in the blue box and appendix in the green box represent the articles that were published during the development of the thesis. The solid arrows indicate the dependencies between the chapters (i.e., one chapter must be read before the other for a better understanding of the proposed techniques). In addition, the dashed arrows indicate the relationship between each chapter and the appendix.

This thesis starts with an overview of face synthesis and data augmentation techniques in the first chapter. They are presented for still-to-video FR applications, which is the main concern examined in this thesis.

Chapter 2 presents a domain-specific face synthesizing technique to improve the performance of still-to-video FR systems when surveillance videos are captured under various uncontrolled conditions, and individuals are recognized based on a single facial image. The proposed approach takes advantage of target domain information from the generic set that can effectively represent probe ROIs. A representative set of synthetic faces is generated that resemble individuals of interest under capture conditions relevant to the target domain. For proof-of-concept

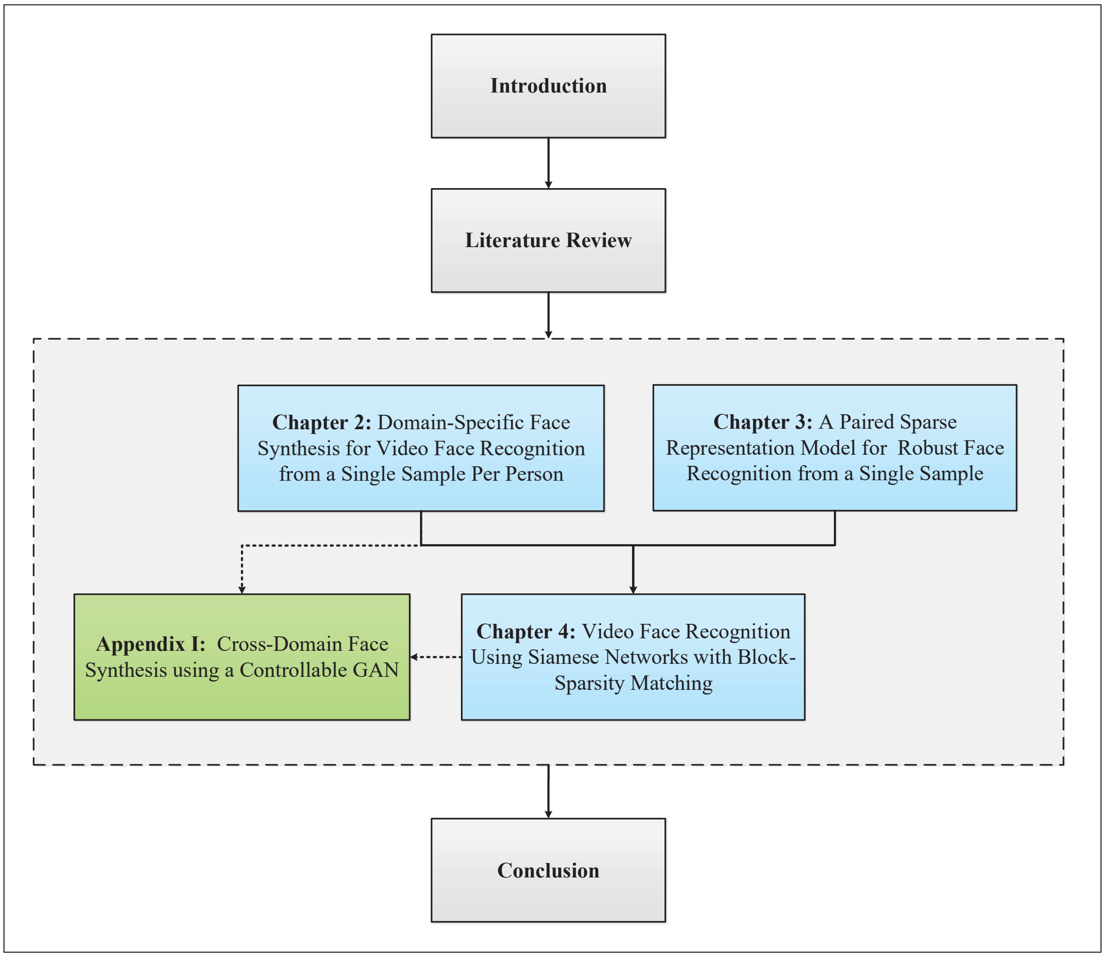

Introduction
Literature Review
Chapter 2: Domain-Specific Face Synthesis for Video Face Recognition from a Single Sample Per Person
Chapter 3: A Paired Sparse Representation Model for Robust Face Recognition from a Single Sample
Appendix I: Cross-Domain Face Synthesis using a Controllable GAN
Chapter 4: Video Face Recognition Using Siamese Networks with Block-Sparsity Matching
Conclusion

Figure 0.1 The flow of the thesis is shown by connected boxes. The solid arrows indicate the dependencies between the chapters and appendix (i.e., one chapter must be read beforehand). Dashed arrows indicates the suggested readings between the chapters and appendix for a better comprehension.

validation, an augmented dictionary with a block structure is designed based on the proposed face synthesizing, and face classification is performed within a sparse representation framework.

In Chapter 3, a paired sparse reconstruction model is presented to account for linear and non-linear variations in the context of still-to-video FR. The proposed model leverages both face synthesis and generic learning to effectively represent probe ROIs from a single reference still. This approach manages non-linear variations by enriching the gallery dictionary with a representative set of synthetic profile faces, where synthetic (still) faces are paired with generic set (video) face in the auxiliary variational dictionary. In this model, the augmented gallery dictionary is encouraged to share the same sparsity pattern with the auxiliary dictionary for the same pose angles. In this way, each synthetic profile image in the augmented gallery dictionary is combined with approximately the same facial viewpoint in the auxiliary dictionary which is a more accurate way of representation, and allows for a higher level of FR accuracy.

Chapter 4 presents an approach that exploits a deep Siamese network and sparse representation-based classification with block structure for pair-wide face matching. It also leverages domain-specific face synthesis to further improvement where rendering parameters are obtained through row sparsity clustering of unlabeled faces captured in the target domain. The proposed technique improves the performance of still-to-video FR systems when surveillance videos are captured under various uncontrolled conditions, and individuals must be recognized based on a single facial still.

Appendix 1 presents a cross-domain face synthesis approach based on a controllable GAN that learns a model using synthetic images as inputs instead of random noise vector. It generates a set of realistic synthetic facial images under the target domain capture conditions with high consistency, while preserving their identity and allowing to specify the pose conditions of synthetic images.

CHAPTER 1

LITERATURE REVIEW

This chapter provides a comprehensive survey and critical analysis of systems for still-to-video FR and state-of-the-art techniques to address their challenges, in particular face synthesis and generic learning.

1.1 Still-to-Video Face Recognition Systems

In a still-to-video FR scenario, there is typically one or more still image(s) to enroll an individual to the system while a set of video frames is available for recognition. Given one or few reference still images, still-to-video FR system seeks to detect the presence of target individuals enrolled to the system over a network of surveillance cameras. In recent years, few specialized approaches have been proposed for still-to-video FR in the literature. Bashbaghi *et al.* (2017b) proposed a robust still-to-video FR system based on multiple face representations. In this work, various feature extraction techniques are applied to face patches isolated in the single reference sample to generate multiple face representations to make it robust to nuisance factors commonly found in video surveillance applications. An individual-specific ensemble of exemplar-SVM classifiers is proposed by Bashbaghi *et al.* (2017a) to develop a domain adaptive still-to-video FR to improve its robustness to intra-class variations. Parchami *et al.* (2017c) developed an accurate still-to-video FR from a SSPP based on deep supervised autoencoder that can represent the divergence between the source and target domains. The autoencoder network is trained using a novel weighted pixel-wise loss function that is specialized for SSPP problems, and allows to reconstruct high-quality canonical ROIs for matching. Parchami *et al.* (2017a) presented an efficient network for still-to-video FR from a single reference still based on cross-correlation matching and triplet-loss optimization that provides discriminant face representations. The matching pipeline exploits a matrix Hadamard product followed by a fully connected layer inspired by adaptive weighted cross-correlation. Parchami *et al.* (2017b) introduced an ensemble of CNNs named HaarNet for still-to-video FR, where a trunk network first extracts features from the global appearance of the facial ROIs. Then, three branch networks effectively embed asymmetrical and complex facial features based on Haar-like features. Dewan *et al.* (2016) exploited an adaptive appearance model tracking for still-to-video FR to gradually learn a track-face-model for each individual appearing in the scene. The models are matched over successive frames against the reference still images of each target individual enrolled to the system, and then matching scores are accumulated over several frames for robust spatio-temporal recognition. Migneault *et al.* (2018) considered adaptive visual trackers for still-to-video FR to regroup faces (based on appearance and temporal coherency) that correspond to the same individual captured along a trajectory, and thereby learn diverse appearance models on-line. Mokhayeri *et al.* (2015) designed a practical still-to-video FR system for video surveillance applications by benefiting from face synthesis. the synthetic images are produced based on camera-specific capture conditions.

1.1.1 Challenges

Despite the recent progress in computer vision and machine learning, designing a robust system for still-to-video FR remains a challenging problem in real-world surveillance applications.

Domain Shift

One key issue is the visual domain shift between faces from the source domain, where reference still images are typically captured under controlled conditions, and those from the target domain, where video frames are captured under uncontrolled conditions with variations in pose, illumination, blurriness, etc. The appearance of faces captured in videos of target domain corresponds to multiple non-stationary data distributions that can differ considerably from faces captured in source domain during enrollment (Bashbaghi *et al.* (2017a)).

Single Sample Per Person Scenario

Another key issue is the limited number of reference stills stored in the gallery. In many surveillance applications (e.g., watch-list screening), only a single reference still per person is available for design, which corresponds to the so-called Single Sample Per Person (SSPP) problem. The performance of still-to-video FR systems can decline significantly due to the limited information available to represent the intra-class variations seen in video frames. Many discriminant subspaces and manifold learning algorithms cannot be directly employed with a SSPP problem. It is also difficult to apply representation-based FR methods such as sparse representation-based classification (SRC) (Wanger *et al.* (2009)) and deep learning methods under a SSPP scenario. Although still faces from the cohort, other non-target persons, and trajectories of video frames from unknown individuals are typically available.

Although designing a robust FR system based on a single sample per person in under surveillance conditions is challenging, several techniques have been recently proposed to address these problems and improve the robustness of still-to-video FR systems designed using a single sample accordingly. They can be categorized into techniques for (1) data augmentation (Masi *et al.* (2019b)), (2) multiple face representation (Bashbaghi *et al.* (2017a)), (3) normalization (Cao *et al.* (2019b)) and (4) Image patching (Zhu *et al.* (2014)). This Thesis mainly focuses on methods that are based on augmenting the reference gallery set through synthetic set either generic set. Figure 1.1 shows the different categories of existing solutions hierarchy.

The following sections give a review of face synthesis and generic learning methods for data augmentation to address visual domain shift and limited samples problems in FR systems.

1.2 Data Augmentation

Generating synthetic face images from a single face image has a wide range of applications in the field of FR. Shekhar *et al.* (2017) enhanced the SRC performance for FR by augmenting the reference gallery using the synthetically relighted images. Masi *et al.* (2016) proposed a data augmentation technique that enriches the training dataset with important facial appearance

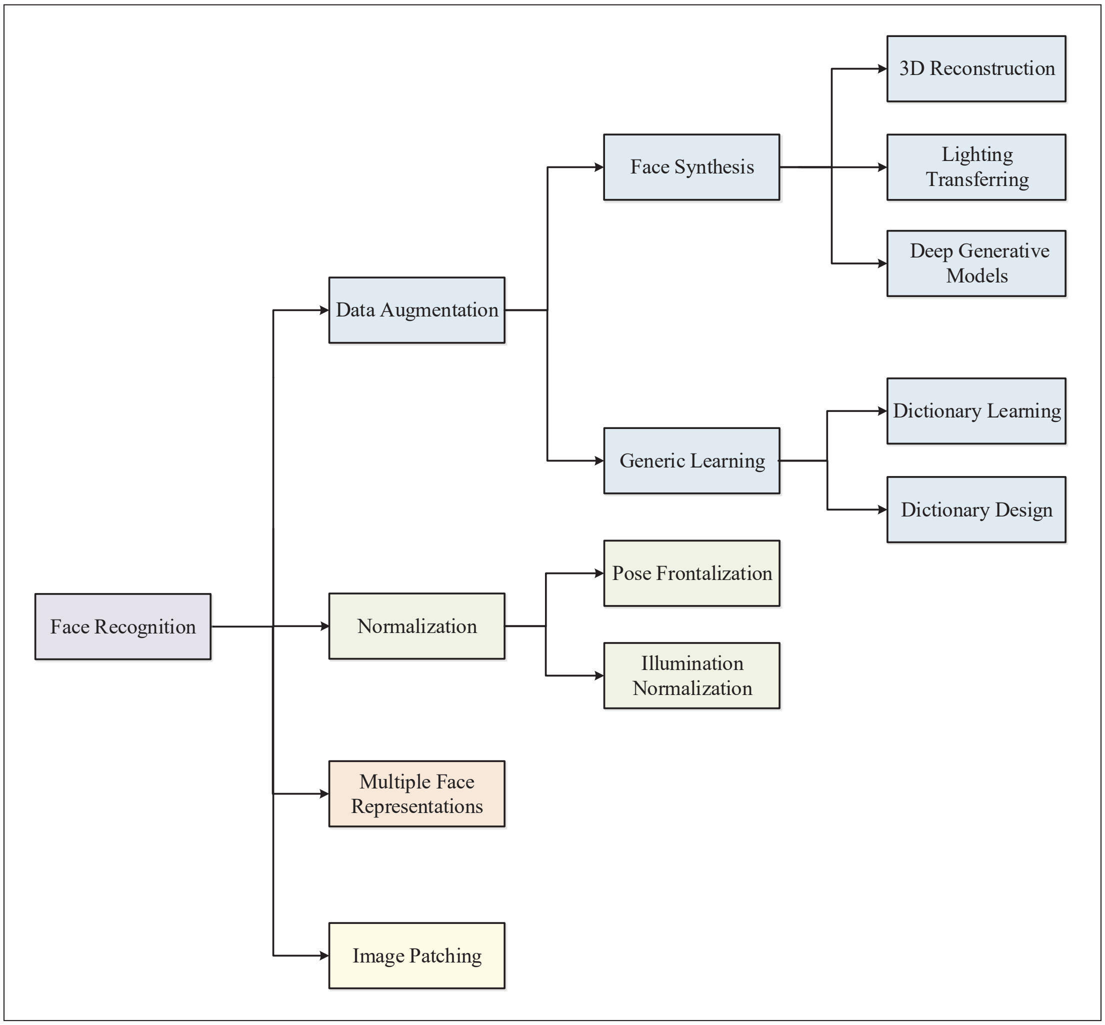


Figure 1.1 Taxonomy of the solutions proposed in the literature to address challenging issues of FR systems.

variations by manipulating the faces it contains. An efficient face-specific data augmentation technique has been introduced by Masi *et al.* (2019b) that uses a fast rendering during training to augment the training set with intra-subject appearance variations, thus effectively training our CNN on a much larger training set.

1.2.1 Face Synthesis

1.2.1.1 3D Morphable Model

A common approach for synthetic face generation is to reconstruct the 3D model of a face using its 2D face image. As a classic statistical model of 3D facial shape and texture, 3D Morphable Model (3DMM) is widely used to reconstruct a 3D face from a single 2D face image and accordingly synthesize new face images (Blanz & Vetter (2003)). This algorithm is based on designing a morphable model from 3D scans and fitting the model to 2D images for 3D shape and texture reconstruction. The 3DMM is based on two key ideas: first, all faces are in dense point-to-point correspondence, which is usually established on a set of example faces in a registration procedure and then maintained throughout any further processing steps. The second idea is to separate facial shape and color and to disentangle these from external factors such as illumination and camera parameters. The Morphable Model may involve a statistical model of the distribution of faces, which was a principal component analysis in the original work and has included other learning techniques in subsequent work.

In the past decade, several extensions of 3DMM is presented for 3D face reconstruction. Zhang & Samaras (2006) proposed a 3D spherical harmonic basis morphable model that is an integration of spherical harmonics into the 3DMM framework. More recently, Koppen *et al.* (2018) expanded 3DMM by adopting a shared covariance structure to mitigate small sample estimation problems associated with data in high dimensional spaces. It models the global population as a mixture of Gaussian sub-populations, each with its own mean value. Gecer *et al.* (2019) revisited the original 3DMM fitting making use of non-linear optimization to find the optimal latent parameters that best reconstruct the test image. They optimized the parameters with the supervision of pre-trained deep identity features through an end-to-end differentiable framework.

Despite the significant success of 3DMM-based techniques for 3D face modeling they often fail to represent small details since they are not spanned by the principal components. An alternative line of work considers CNNs for 3D face modeling with 3DMM. Embedding 3D morphable basis functions into deep neural networks opens great potential for models with better representation power which is superior in capturing a higher level of details. Tran *et al.* (2019) improved the nonlinear 3DMM in both learning objective and architecture by solving the conflicting objective problem with learning shape and albedo proxies with proper regularization. The novel pairing scheme allows learning both detailed shape and albedo without sacrificing one. Tran *et al.* (2017a) employed a CNN to regress 3DMM shape and texture parameters directly from an input image without an optimization process which renders the face and compares it to the image. Richardson *et al.* (2017) presented a face reconstruction technique from a single image by introducing an end-to-end CNN framework which derives a novel rendering layer, allowing back-propagation from a rendered depth map to the 3DMM model. In the same line, Tewari *et al.* (2017) proposed a CNN regression-based approach for face reconstruction, where a single forward pass of the network estimates a much more complete face model, including pose, shape, expression, and illumination, at a high quality. Due to the type and amount of training data, as well as, the linear bases, the representation power of 3DMM can be limited. To address these problems, Tran & Liu (2018) proposed an innovative framework to learn a nonlinear 3DMM model from a large set of in-the-wild face images, without collecting 3D face scans. Specifically, given a face image as input, a network encoder estimates the projection, lighting, shape and albedo parameters. Two decoders serve as the nonlinear 3DMM to map from the shape and albedo parameters to the 3D shape and albedo, respectively.

Although their results are encouraging, the synthetic face images may not be realistic enough to represent intra-class variations of target domain capture conditions. The synthetic images generated in this way are highly correlated with the original facial stills from enrolment, and there is typically a domain shift between the distribution of synthetic faces and that of faces captured in the target domain which poses the problem of domain adaptation. The FR models naively trained on these synthetic images, often fail to generalize well when matched to real images captures in the target domain. Producing realistic synthetic face images while preserving their identity information is still an ill-posed problem.

1.2.1.2 Generative Adversarial Network

Recently, Generative Adversarial Networks (GANs) have shown promising performance in generating realistic images (Gonzalez-Garcia *et al.* (2018); Choi *et al.* (2018); Chen & Koltun (2017)). GANs are a framework to produce a model distribution that mimics a given target distribution, and it consists of a generator that produces the model distribution and a discriminator that distinguishes the model distribution from the target. The concept is to consecutively train the model distribution and the discriminator in turn, with the goal of reducing the difference between the model distribution and the target distribution measured by the best discriminator possible at each step of the training (Goodfellow & et al. (2014)). Benefiting from GAN, FaceID-GAN is proposed by Shen *et al.* (2018) which generates photorealistic and identity preserving faces. It competes with the generator by distinguishing the identities of the real and synthesized faces to preserve the identity of original images. Gecer *et al.* (2018) proposed a novel end-to-end semi-supervised adversarial framework to generate photorealistic face images of new identities with a wide range of expressions, poses, and illuminations conditioned by synthetic images sampled from a 3DMM. Huang *et al.* (2017a) proposed TP-GAN for photorealistic frontal view synthesis by simultaneously perceiving global structures and local details. They made problem well constrained by introducing a combination of adversarial loss, symmetry loss and identity preserving loss. The combined loss function leverages both frontal face distribution and pre-trained discriminative deep face models to guide an identity preserving inference of frontal views from profiles. Wang *et al.* (2018b) proposed a variant of GANs for face aging in which a conditional GAN module functions as generating a face that looks realistic and is with the target age, an identity-preserved module preserves the identity information and an age classifier forces the generated face with the target age. Tewari *et al.* (2017) proposed a novel model-based deep convolutional autoencoder for 3D face reconstruction from a single in-the-wild color image that combine a convolutional encoder network with a model-based face reconstruction model. In this way, the CNN-based encoder learns to extract semantically meaningful parameters from a single monocular input image. WGAN is a recent technique which employs integral probability metrics based on the earth mover distance rather than the Jensen–Shannon divergence that the original GAN uses (Arjovsky *et al.* (2017)). BEGAN built upon WGAN using an autoencoder based equilibrium enforcing technique alongside the Wasserstein distance to stabilize the training of the discriminator (Berthelot *et al.* (2017)).

Difficulty in controlling the output of the generator is a challenging issue in GAN-based face synthesis models. To reduce this gap, conditional GANs are proposed that leverage conditional information in the generative and discriminative networks for conditional image synthesis. Tran *et al.* (2018) used pose codes in conjunction with random noise vectors as the inputs to the discriminator with the goal of generating a face of the same identity with the target pose in order to fool the discriminator. Hu *et al.* (2018) introduced a coupled-agent discriminator which forms a mask image to guide the generator during the learning process. Mokhayeri *et al.* (2019b) proposed a controllable GAN that employs an additional adversarial game as the third player to the GAN, competing with the generator to preserve the specific attributes, and accordingly, providing control over the face generation process. Despite the success of GAN in generating realistic images, they still struggle in learning complex underlying modalities in a given dataset, resulting in poor-quality generated images.

1.2.2 Generic Learning

Generic learning is another effective solution to compensate visual domain shift in FR systems that employs a generic set to enrich the diversity of the reference gallery set (Gao *et al.* (2017); Li *et al.* (2016)). Su *et al.* (2010) proposed an adaptive generic learning method for FR which utilized external data to estimate the within-class scatter matrix for each individual and applies this information to the reference set. Recent reports have suggested that integration of SRC with generic learning substantially boosts the performance of FR systems. Deng *et al.* (2012) added generic learning into the SRC framework and proposed the extended SRC (ESRC), which provide additional information from other face datasets to construct an intra-class variation dictionary to represent the changes between the training and probe images. Yang *et al.* (2013) introduced a sparse variation dictionary learning (SVDL) technique by taking the relationship between the reference set and the external generic set into account and obtained a projection by learning from both generic and reference set. Nourbakhsh *et al.* (2016) integrated intra-class variation information from the target domain with the reference set through domain adaptation to enhance the facial models. Wei & Wang (2015) designed a robust auxiliary dictionary learning technique that extracts representative information from generic dataset via dictionary learning without assuming prior knowledge of occlusion in probe images. Zhu *et al.* (2014) proposed a local generic representation-based framework for FR with SSPP. It builds a gallery dictionary by extracting the neighboring patches from the gallery dataset, while an intra-class variation dictionary is constructed by using an external generic training dataset to predict the intra-class variations. Bashbaghi *et al.* (2015) proposed a robust still-to-video FR using a multi-classifier system in which each classifier is trained by a reference face still versus many lower-quality faces of non-target individuals captured in videos. In this system, the auxiliary set collected from the videos of unknown people in the target domain is employed to select discriminant feature sets and ensemble fusion functions. Despite the significant improvements reported with generic learning, several critical issues remain to be addressed. The generic intra-class variation may not be similar to that of gallery individuals, so the extraction of discriminative information from the generic set may not be guaranteed.

1.3 Deep Face Recognition

Recent state-of-the-art approaches for video FR mostly rely on CNNs (Parchami *et al.* (2017c); Cao *et al.* (2018); Zhao *et al.* (2018); Parkhi *et al.* (2015)). One of the pioneering work in this domain is Siamese network (Bromley *et al.* (1994)) that formulates deep learning with a contrastive loss that minimizes distance between positive pairs while keeps negative pairs apart. Deep Siamese networks are often designed using two or more identical sub-networks and one cost module where the extractors share same parameters and weights. During the training process, these networks typically seek to minimize the intra-class distance and maximize the inter-class distances. When the features are extracted for a pair of images, the matcher produces a similarity score indicating if the pair images are from the same or different classes.

Siamese networks have significantly improved FR accuracy due to their high capacity for learning discriminative features (Parchami *et al.* (2017a)). Taigman *et al.* (2014) proposed to employ these networks to learn similarity metrics for FR which trained on a large dataset from Facebook. Schroff *et al.* (2015) introduced FaceNet that directly learns a mapping from face images to a compact Euclidean space. FaceNet uses a deep Siamese network that directly optimizes the L_1-distance between two faces and employs face triplets and minimizes the distance between an anchor and a positive sample of the same identity, while maximizing the distance between the anchor and a negative sample of a different identity. Light CNN framework was proposed by Wu *et al.* (2018a) to learn deep face representations from the large-scale dataset with noisy labels, where a max-feature map operation allows to obtain a compact representation. Yin & Liu (2017) presented a multi-task CNN for FR that exploits side tasks in regularization to learn pose-specific identity features. Masi *et al.* (2019a) proposed a pose-aware network to process a face image using several pose-specific CNNs. Parchami *et al.* (2017b) introduced an ensemble of CNNs named HaarNet for FR, where a trunk network first extracts features from the global appearance of the facial ROIs. Then, three branch networks effectively embed asymmetrical and complex facial features based on Haar-like features. Peng *et al.* (2019) developed a deep local descriptor for cross-modality FR, which can learn discriminant information from image patches. Despite the great success of CNNs in FR (Schroff *et al.* (2015)), contrastive embedding requires training data contains real-valued precise pair-wise similarities or distances. This problem is solved by optimizing the relative distance of the positive pair and one negative pair from three samples (Salakhutdinov & Hinton (2007)). To facilitate the training process, the N-pair loss (Sohn (2016)) is introduced to consider multiple negative samples in training, and exhibits higher efficiency and performance. More recently, the angular loss is proposed by Wang *et al.* (2017) to enhance N-pair loss by integrating high-order constraint that captures additional local structure of triplet triangles.

1.4 Summary

Reviewing the current literature on video FR, we identify that designing a robust face FR under surveillance conditions is still a challenging task, and the performance of such systems declines when the number of training images per class is limited and the underlying distribution between the still reference and probe video ROIs differ. Deep learning and representation-based FR methods such as sparse representation-based classification cannot be directly employed with limited samples. Extending the gallery using the synthetic face images to address these problems is the main topic of this thesis, which is explored in chapters 2, 3 and 4.

Generating photorealistic synthetic face images being able to cover target domain variations with high consistency is still a challenge. We address this problem by integrating an image-based face relighting technique inside the 3D reconstruction framework and projecting the discriminant information of the generic set onto the reference stills. Finding representative variations to prevent over-fitting is another key aspect of data augmentation. In this thesis, we select representative facial exemplars by applying clustering on the information extracted from the videos of target domain. Another key challenge is preserving the identity information of the generated faces, which is critical for FR applications. In this thesis, we address this issue by presenting an extended GAN conditioned by synthetic images that uses an additional adversarial game as the third player to the original GAN, competing with the generator to preserve the specific attributes.

CHAPTER 2

DOMAIN-SPECIFIC FACE SYNTHESIS FOR VIDEO FACE RECOGNITION FROM A SINGLE SAMPLE PER PERSON

Fania Mokhayeri[1], Eric Granger[1], Guillaume-Alexandre Bilodeau[2]

[1] Le Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA),
École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Laboratoire d'Interprétation et de Traitement d'Images et Vidéo (LITIV),
Polytechnique Montréal,
Montréal, Québec, Canada H3T 1J4

Article published in « IEEE Transactions on Information Forensics and Security » 2018.

Abstract

In video surveillance, face recognition (FR) systems are employed to detect individuals of interest appearing over a distributed network of cameras. The performance of still-to-video FR systems can decline significantly because faces captured in unconstrained target domain over multiple video cameras have a different underlying data distribution compared to faces captured under controlled conditions in the source domain with a still camera. This is particularly true when individuals are enrolled to the system using a single reference still. To improve the robustness of these systems, it is possible to augment the reference set by generating synthetic faces based on the original still. However, without knowledge of the target domain, many synthetic images must be generated to account for all possible capture conditions. FR systems may, therefore, require complex implementations and yield lower accuracy when training on many less relevant images. This paper introduces an algorithm for domain-specific face synthesis (DSFS) that exploits the representative intra-class variation information available from the target domain. Prior to operation (during camera calibration), a compact set of faces from unknown persons appearing in the target domain is selected through affinity propagation clustering in the captured condition space (defined by pose and illumination estimation). The domain-specific variations of these face images are then projected onto the reference still of

each individual by integrating an image-based face relighting technique inside the 3D reconstruction framework. A compact set of synthetic faces is generated that resemble individuals of interest under the capture conditions relevant to the target domain. In a particular implementation based on sparse representation classification, the synthetic faces generated with the DSFS are employed to form a cross-domain dictionary that accounts for structured sparsity where the dictionary blocks combine the original and synthetic faces of each individual. Experimental results obtained with videos from the Chokepoint and COX-S2V datasets reveal that augmenting the reference gallery set of still-to-video FR systems using the proposed DSFS approach can provide a significantly higher level of accuracy compared to state-of-the-art approaches, with only a moderate increase in its computational complexity.

2.1 Introduction

Still-to-video face recognition (FR) is an important function in several video surveillance applications, particularly for watch-list screening. Given one or more reference still images of a target individual of interest, still-to-video FR systems seeks to accurately detect their presence in videos captured over multiple distributed surveillance cameras (Dewan *et al.* (2016)).

Despite the recent progress in computer vision and machine learning, designing a robust system for still-to-video FR remains a challenging problem in real-world surveillance applications. One key issue is the visual domain shift between faces from the source domain (enrollment domain), where reference still images are typically captured under controlled conditions, and those from the target domain (operational domain), where video frames are captured under uncontrolled conditions with variations in pose, illumination, blurriness, etc. The appearance of faces captured in videos corresponds to multiple non-stationary data distributions that can differ considerably from faces captured during enrollment (Bashbaghi *et al.* (2017a)). Another key issue is the limited number of reference stills that are available per target individual to design facial models. Although still faces from the cohort or other non-target persons, and trajectories of video frames from unknown individuals are typically available. In many surveillance applications (e.g., watch-list screening), only a single reference still per person is available for design, which corresponds to the so-called Single Sample Per Person (SSPP) problem. The performance of still-to-video FR systems can decline significantly due to the limited information available to represent the intra-class variations seen in video frames. Many discriminant subspaces and manifold learning algorithms cannot be directly employed with a SSPP problem. It is also difficult to apply representation-based FR methods such as sparse representation-based classification (SRC) (Wright *et al.* (2009)).

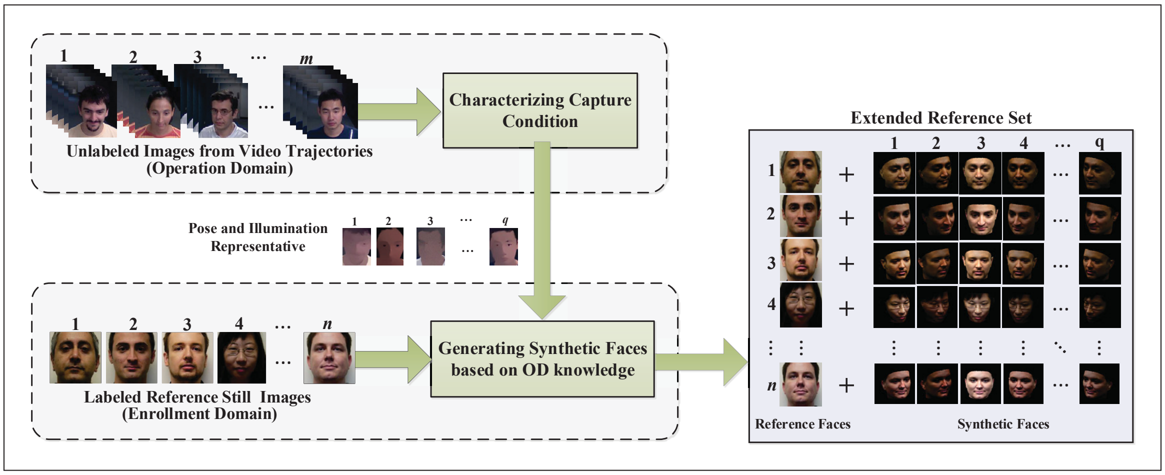

1
2
3
...
m
Unlabeled Images from Video Trajectories
(Operation Domain)
Characterizing Capture Condition
Pose and Illumination Representative
1
2
3
...
q
1
2
3
4
...
n
Labeled Reference Still Images
(Enrollment Domain)
Generating Synthetic Faces based on OD knowledge
Extended Reference Set
1
2
3
4
...
q
+
Reference Faces
Synthetic Faces

Figure 2.1 Overview of the proposed DSFS algorithm to augment the reference gallery set. We assume that the gallery set initially contains only one reference still image per individual of interest.

Different techniques for SSPP problems have been proposed to improve the robustness of FR systems, such as using multiple face representations (Bashbaghi *et al.* (2017a)), face frontalization (Hassner *et al.* (2015)), generating synthetic faces from the original reference stills (Liu & Wassell (2015); Masi *et al.* (2016)), and incorporating generic auxiliary set (Deng *et al.* (2012, 2014)). This paper focuses on methods that are based on augmenting the gallery set using synthetic set generated based on the original reference still, and by taking into account the generic variational information. A challenge with strategies for augmenting the reference gallery set is selecting a sufficient number of synthetic or generic faces to cover intra-class variations in the target domain. Many synthetic faces or generic auxiliary faces may be generated or collected, respectively, to account for all possible capture conditions. FR systems would, therefore, require complex implementations and may yield lower accuracy when training on many facial images. Another challenge is domain discrepancy between synthetic and real images. The synthetically generated images may not be covering the range intra-class variations of target domain, since they are highly correlated with the original face images.

In this paper, a new approach is proposed that exploits the discriminant information of the generic set for the face synthesis process. The new algorithm called domain-specific face synthesis (DSFS) maps representative variation information from the generic set in the target domain to the original reference stills. In this way, a compact set of synthetic faces is generated that represent reference still images and probe video frames under a common capture condition. As depicted in Fig. 2.1, the DSFS technique involves two main steps: (1) characterizing capture condition information from the target domain, (2) generating synthetic face images based on the information obtained in the first step. Prior to operation (during camera calibration process), a generic set is collected from video captured in the target domain. A compact and representative subset of face images is selected by clustering this generic set in a capture condition space defined by pose, illumination, blur. The 3D model of each reference still image is reconstructed via a 3D morphable model and rendered based on pose representatives. Finally, the illumination-dependent layers of the lighting representatives are extracted and projected on the rendered reference images with the same pose. In this manner, domain-specific variations are effectively transferred onto the reference still images. The major contributions of our work are:

- A technique based on affinity clustering to select representative facial exemplars using information extracted from videos captured form target domain. This prevents over-fitting of classifier due to the redundant information and improves efficiency.
- A novel face synthesizing technique that maps the intra-class variation from facial exemplars available in the target domain to generate a representative set of face images under real-world capture conditions.
- A technique to design a compact and discriminative dictionary for SRC allowing to perform robust still-to-video FR with only one reference still ROI.

In a particular implementation for still-to-video FR, the original and synthetic face images are employed to design a structural dictionary with powerful variation representation ability for SRC. The dictionary blocks represent intra-class variations computed from either the reference faces themselves or the synthetic faces (Elhamifar & Vidal (2011)). The cooperation of SRC with the proposed DSFS improves the robustness of SRC for video-based FR in a SSPP scenario to domain variations. In order to validate the performance of the proposed DSFS algorithm for still-to-video FR with a SSPP, this SRC implementation is evaluated and compared on two public face databases.

The main advantage of the proposed approach is the ability to provide a compact set that can accurately represent the original reference face with relevant of intra-class variations in pose, illumination, motion blur, etc., corresponding to capture condition in the target domain. For instance, in the context of SRC implementations, this set can prevent over-fitting and refines more informative classes during the sparse coding process. The rest of the paper is organized as follows. Section 2.2 provides an overview of related works for FR with a SSPP. Section 2.3 describes the proposed face synthesizing algorithm. Section 2.4 presents a particular implementation of the DSFS for still-to-video FR system. In Section 2.5, the experimental methodology (dataset, protocol, and performance metrics) for validation of FR systems is described, and the experimental results is presented in 2.6. Finally, Section 2.7 concludes the paper and discusses some future research directions.

2.2 Related Work

Several techniques have been proposed in the literature to improve the robustness of still-to-video FR systems designed using a SSPP. They can be categorized into techniques for (1) multiple face representation, (2) generic learning, and (3) generation of synthetic faces. An overview of the techniques is presented as below.

2.2.1 Multiple Face Representations

One effective approach to address the SSPP problem in FR is to extract discriminant features from face images. Bashbaghi *et al.* (2014) developed a robust still-to-video FR system based on diverse face representations. They applied multiple appearance-invariant feature extraction techniques to patches isolated from the reference still images in order to produce multiple face representations and generate a pool of diverse exemplar-SVMs. This pool provides robustness to common nuisance factors encountered in surveillance applications. Lu *et al.* (2013) proposed a discriminative multi-manifold analysis method by learning discriminative features from image patches. In this technique, the patches of each individual are considered to form a manifold for each sample per person and a projection matrix is learned by maximizing the manifold margin of different persons. A deep face representation method is proposed by Almageed *et al.* (2016) using several pose-specific deep CNN models to generate multiple pose-specific features. The multiple face representation techniques are, however, able to compensate only the small variations and consequently are not effective to tackle with variations in practical applications (e.g., extreme illumination, pose and expression variations).

2.2.2 Generic Learning

An early finding to compensate visual domain shift in FR systems is to employ a generic set to enrich the diversity of the reference gallery set that is the so-called generic learning concept (Su *et al.* (2010)). Generic learning has been widely discussed by many researchers (Gao *et al.* (2017); Li *et al.* (2016)). Su *et al.* (2010) proposed an adaptive generic learning method for FR which utilized external data to estimate the within-class scatter matrix for each individual and applies this information to the reference set. In recent years, integration of sparse representation-based classification (SRC) with generic learning for FR has attracted significant attention. Deng *et al.* (2012) added generic learning into the SRC framework and proposed the extended SRC (ESRC), which provide additional information from other face datasets to construct an intra-class variation dictionary to represent the changes between the training and probe images. Yang *et al.* (2013) introduced a sparse variation dictionary learning (SVDL) technique by taking the relationship between the reference set and the external generic set into account and obtained a projection by learning from both generic and reference set. Nourbakhsh *et al.* (2016) proposed integrating intra-class variation information from the target domain with the reference set through domain adaptation to enhance the facial models. Wei & Wang (2015) proposed a robust auxiliary dictionary learning (RADL) technique that extracts representative information from generic dataset via dictionary learning without assuming prior knowledge of occlusion in probe images. Zhu *et al.* (2014) proposed a local generic representation-based framework (LGR) for FR with SSPP. It builds a gallery dictionary by extracting the neighboring patches from the gallery dataset, while an intra-class variation dictionary is constructed by using an external generic training dataset to predict the intra-class variations. A supervised autoencoder network for still-to-video FR system is proposed by Parchami *et al.* (2017c) that generates canonical face representations from unknown video frames in the target domain that are robust to appearance variations. Despite the significant improvements reported with generic learning, several critical issues remain to be addressed. The generic intra-class variation may not be similar to that of gallery individuals, so the extraction of discriminative information from the generic set may not be guaranteed. Moreover, the large number of images collected from external data may contain redundant information which could lead to complex implementations and degrade the capability in covering intra-class variations.

2.2.3 Synthetic Face Generation

Augmenting the reference gallery set synthetically is another strategy to compensate the appearance variations in FR with SSPP. Shao *et al.* (2017) presented a SRC-based FR algorithm that extends the dictionary using a set of synthetic faces generated by calculating the image difference of a pair of faces. Mokhayeri *et al.* (2015) augmented the reference gallery set by generating a set of synthetic face images under camera-specific lighting conditions to design a robust still-to-video FR system under surveillance conditions. 3D Morphable Model (3DMM), proposed by Blanz & Vetter (2003), has been widely used to synthesize new face images from a single 2D face image. Tran *et al.* (2017a) employed a CNN to regress 3DMM shape and texture parameters directly from an input image without an optimization process which renders the face and compares it to the image. Zhang & Samaras (2006) proposed a 3D Spherical Harmonic Basis Morphable Model (SHBMM) that is an integration of spherical harmonics into the 3DMM framework. Richardson *et al.* (2017) proposed a neural network for reconstructing a detailed facial surface in 3D from a single image where the rough facial geometries are modeled using a 3DMM and facial features of that geometry is refined by a CNN. The proposed method by Tewari *et al.* (2017) integrates an expert-designed decode layer that implements an elaborate generative analytically-differentiable image formation model on the basis of a detailed parametric 3D face model. Apart from 3D reconstruction techniques, some 2D-based techniques generate synthetic images under various illumination conditions by transferring the illumination of target images to the reference face images (Isola *et al.* (2017)). Liu *et al.* (2019) proposed a encoder-decoder framework that for the first time jointly learns face models directly from raw scans of multiple 3D face databases and establishes dense correspondences among all scans. Recently, generative adversarial network (GAN), introduced by Goodfellow & et al. (2014) has become popular for realistic face synthesis (Shen *et al.* (2018); Tran *et al.* (2017b); Bao *et al.* (2018b)). These methods formulate GAN as a minimax game, where a discriminator distinguishes face images in the real and synthetic domains, while a generator reduces its discriminativeness by synthesizing realistic face images. The competition converges when the discriminator is unable to differentiate between real and synthetic domains. Shrivastava *et al.* (2017) proposed Simulated+Unsupervised learning method that improves the realism of synthetic images. The proposed learning method employs an adversarial network similar to GAN with synthetic images as inputs instead of random vectors. Although synthetic images can improve the robustness of FR systems designed with a SSPP, they may not be covering the range intra-class variations in practical scenarios because of redundancy in the learned discriminative subspace. Many synthetic images should be generated to account for all possible capture conditions in target domain. Without the selection of representative face images from both the reference gallery and external data, generating the synthetic faces may require complex implementations and yield lower accuracy when training on many less relevant images.

To overcome the challenges discussed above, this paper presents a framework that exploits both face synthesis and generic learning. The technique proposed in Section III generates a compact set of synthetic facial images per individual of interest that corresponds to relevant target domain capture conditions, by mapping the intra-class variations from a representative set of video frames selected from the target domain into the original reference still images.

2.3 Domain-Specific Face Synthesis

This paper focuses on augmenting reference face set to cover the intra-class variations of individual appearing in target domain with a compact set of synthetic face images. A new Domain-Specific Face Synthesis (DSFS) technique is proposed that employs knowledge of the target domain to generate a compact set of synthetic face images for the design of FR systems. Prior to operation, e.g., during a camera calibration process, DSFS selects facial regions of interest (ROIs) isolated in videos with representative pose angles and illumination conditions from facial trajectories of unknown persons captured in the target domain. These video ROIs are selected via clustering facial trajectories in the captured condition space defined by pose and illumination conditions. Next, the DSFS exploits a 3D shape reconstruction method and an image-based illumination transferring technique to generate synthetic ROIs under representative pose angles and illumination conditions from the reference still ROIs. To do so, the 3D models of the reference still ROIs are reconstructed and rendered w.r.t. the representative pose angles. The illumination-dependent layers of the representative illumination conditions are then extracted and projected onto rendered images with the same view by applying a morphing between the layers. In other words, illumination-dependent layers of video ROIs from the target domain are replaced with that of the still reference ROI from the source domain. Fig. 2.2 shows the pipeline of the DSFS technique.

2.3.1 Characterizing the Capture Conditions

An important concern for the reference set augmentation is the selection of representative pose angles and illumination conditions to represent relevant capture conditions in the target domain.

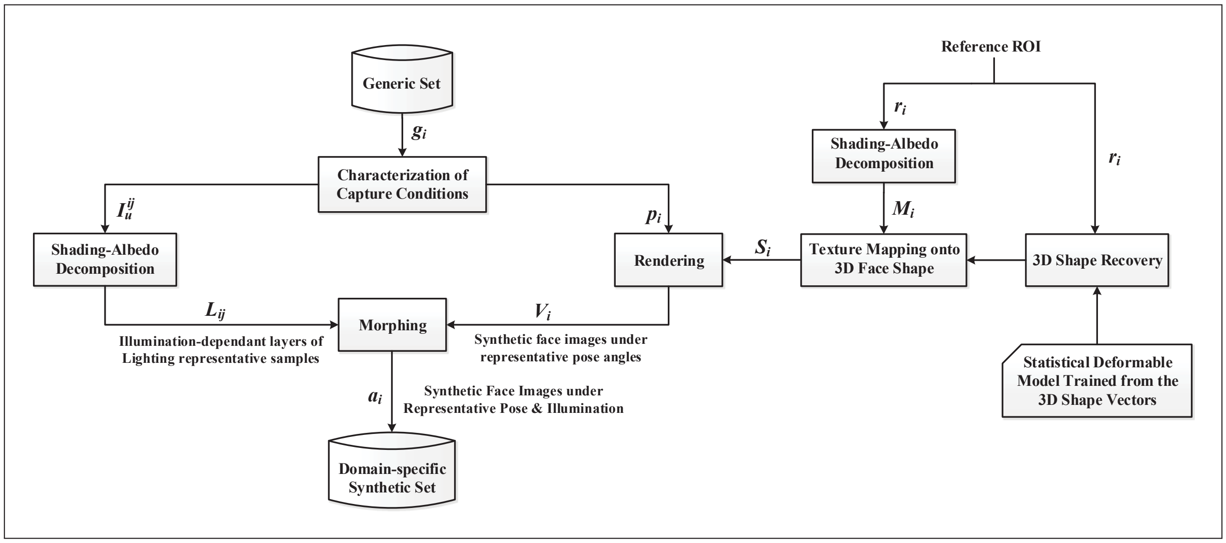

Generic Set
g_i
Characterization of Capture Conditions
I_u^{ij}
p_i
Shading-Albedo Decomposition
Rendering
L_{ij}
Morphing
V_i
Illumination-dependant layers of Lighting representative samples
Synthetic face images under representative pose angles
a_i
Synthetic Face Images under Representative Pose & Illumination
Domain-specific Synthetic Set
Reference ROI
r_i
Shading-Albedo Decomposition
r_i
M_i
S_i
Texture Mapping onto 3D Face Shape
3D Shape Recovery
Statistical Deformable Model Trained from the 3D Shape Vectors

Figure 2.2 Block diagram of the DSFS technique applied to a reference still ROI.

As mentioned, adding a large number of potentially redundant images to the reference set can significantly increase the time and memory complexity, and may degrade the recognition performance due to over-fitting.

With the DSFS technique, the representative pose angles and illumination conditions to cover relevant intra-class variations is approximated by characterizing the capture conditions from a large generic set of video ROIs. This set is formed with multiple ROIs isolated in several facial trajectories of unknown persons captured in the target domain. Let $\mathbf{R} = \{\mathbf{r}_i \in \mathbb{R}^{d \times d} | i = 1, \ldots, n\}$ be a set of ROIs of still reference individuals, and $\mathbf{G} = \{\mathbf{g}_i \in \mathbb{R}^{d \times d} | i = 1, \ldots, m\}$ be a set of video ROIs in the generic set, where n and m denote the number of individuals in the reference gallery set, and the generic set, respectively.

In the proposed technique (see Fig. 2.3), an estimation of luminance, contrast and pose are measured for each video ROIs in the generic set $\mathbf{g}_i$. Next, a two-step clustering process is applied on video ROIs in the measurement space defined by pose, luminance and contrast. The first step is applied on all ROIs in the 3D metric space defined by pose (tilt, yaw and roll), while the second step is applied on ROIs of each pose cluster in the 2D space defined by luminance and contrast metrics. The prototype of each cluster is considered as an exemplar.

The generic variational information obtained during this step is then transferred to the reference still ROIs during the face synthesizing step (see Section III B). Although many algorithms are also suitable to implement DSFS, the following subsections describe DSFS with specific algorithms.

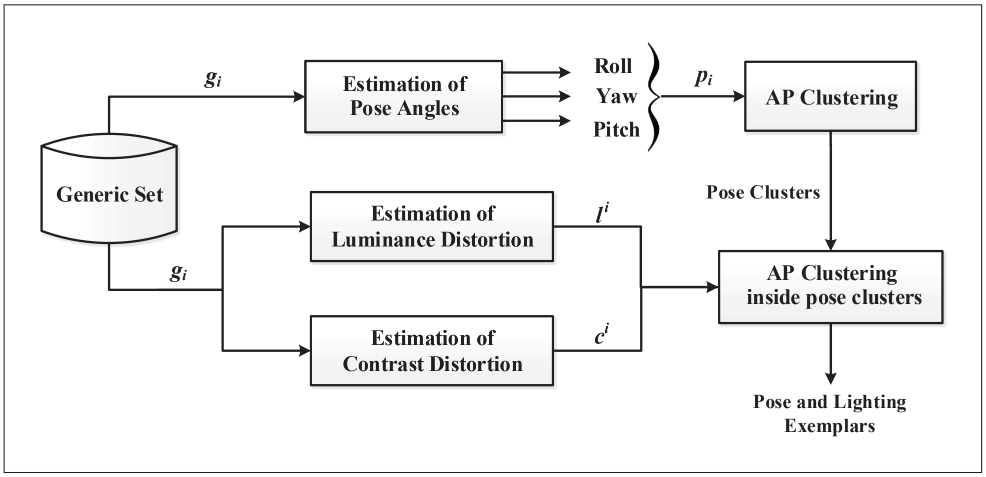

Generic Set
g_i
Estimation of Pose Angles
Roll
Yaw
Pitch
p_i
AP Clustering
Pose Clusters
g_i
Estimation of Luminance Distortion
l^i
Estimation of Contrast Distortion
c^i
AP Clustering inside pose clusters
Pose and Lighting Exemplars

Figure 2.3 Pipeline for characterizing capture conditions of video ROIs in the target domain.

2.3.1.1 Estimation of Head Pose

The estimate of head pose for the i^{th} video ROI ($\mathbf{g}_i$) in the generic set is defined as $\mathbf{p}_i = (\theta_i^{pitch}, \theta_i^{yaw}, \theta_i^{roll})$. Euler angles θ_i^{pitch}, θ_i^{yaw}, and θ_i^{roll} are used to represent pitch, yaw and roll rotation around X axis, Y axis, and Z axis of the global coordinate system, respectively. In order to estimate the head pose, the discriminative response map fitting (DRMF) method is employed (Asthana *et al.* (2013)). It is the current state-of-the-art method in terms of fitting accuracy and efficiency suitable for handling occlusions and changing illumination conditions.

2.3.1.2 Luminance-Contrast Distortion

Luminance and contrast distortion measures estimate the distortion between a video ROI and the corresponding reference still ROI. Components of the structural similarity index measure presented by Wang *et al.* (2004) are employed to measure the proximity of the average luminance and contrast locally by utilizing sliding window. The global luminance distortion in image quality (GLQ) factor between $\mathbf{r}_i$ and $\mathbf{g}_j$ is calculated by sliding a window of $B \times B$ pixels from the top-left corner to the bottom-right corner of the image, for a total of M sliding steps:

$$l^{i,j} = \frac{1}{M} \sum_{\psi=1}^{M} \frac{2.\mu_{\psi}(\mathbf{r}_i).\mu_{\psi}(\mathbf{g}_j) + C_l}{\mu_{\psi}(\mathbf{r}_i)^2 + \mu_{\psi}(\mathbf{g}_j)^2 + C_l} \ , \tag{2.1}$$

where ψ is the sliding step and $\mu_{\psi}(\cdot)$ denotes mean values of the ψ^{th} image window. C_l is a positive stabilizing constant defined as $C_l = (\psi_l L)^2$ where L is the dynamic range of the pixel values and $\psi_l \ll 1$ is a small constant. Similarly, the contrast distortion between $\mathbf{r}_i$ and $\mathbf{g}_i$ is estimated using global contrast distortion in image quality (GCQ) factor defined as:

$$c^{i,j} = \frac{1}{M} \sum_{\psi=1}^{M} \frac{2.\sigma_{\psi}(\mathbf{r}_i).\sigma_{\psi}(\mathbf{g}_j) + C_c}{\sigma_{\psi}(\mathbf{r}_i)^2 + \sigma_{\psi}(\mathbf{g}_j)^2 + C_c} \ , \tag{2.2}$$

where $\sigma_{\psi}(\cdot)$ denotes the standard deviation of the ψ^{th} image window. C_c is a positive stabilizing constant defined as $C_c = (\psi_c L)^2$ where L is the dynamic range of the pixel values and $\psi_c \ll 1$ is a small constant.

2.3.1.3 Representative Selection

Affinity propagation (AP) is applied to cluster video ROIs from the generic set defined in the normalized space defined by $\mathbf{p}_j = (\theta_i^{pitch}, \theta_i^{yaw}, \theta_i^{roll})$ and $\mathbf{u}^i = (l^i, c^i)$ measures (Frey & Dueck (2007)). This clustering algorithm aims to maximize the net similarity (average distortion between ROIs and pose angles) and produce a set of exemplars. Two types of messages: *responsibility* and *availability* are exchanged between data points until a high-quality set of exemplars and corresponding clusters emerges. AP is a suitable clustering technique for DSFS because: (1) it can automatically determine the number of clusters based on the data distribution, and (2) it produced exemplars that correspond to actual ROIs. Indeed, cluster centroids produced by many prototype-based clustering methods are not necessarily actual ROIs with a real-world interpretation. Given that clustering samples simultaneously in terms of both $\mathbf{p}_j$ and $\mathbf{u}^i$ may favor certain common pose angles, a two-step clustering algorithm is proposed to preserves diversity in pose angles and illumination effects. In the first step, clustering is performed on the pose angle vector, and then the population of each pose cluster is clustered according to GLQ and GCQ metrics to find the representative luminance and contrast samples. Representative luminance and contrast samples – called *lighting exemplar* – are found along with representative pose angles – called "*pose exemplar*" (Fig. 2.4).

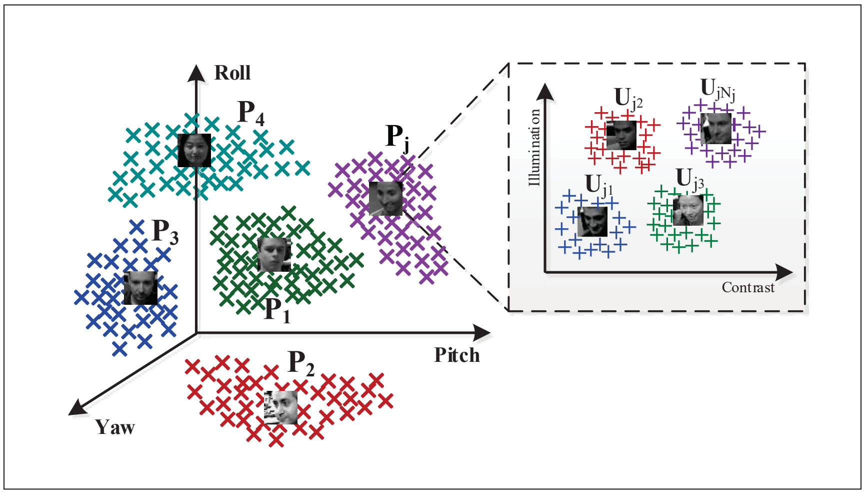


Figure 2.4   An illustration of the AP clustering process.

The clustering algorithm inputs a set of pose similarities $s_p(i,k) = - \parallel \mathbf{p}_i - \mathbf{p}_k \parallel^2$ indicating how well the sample $\mathbf{p}_k$ with index k is similar to the sample $\mathbf{p}_i$ from the generic set. The pose responsibility $r_p(i,k)$ is defined as the accumulated evidence for how well-suited sample $\mathbf{p}_k$ is to serve as the exemplar for the sample $\mathbf{p}_i$, taking into account other potential exemplars for the sample $\mathbf{p}_i$. Evidence about whether each pose candidate exemplar would be a good exemplar is obtained from the application of the pose availability $a_p(i,k)$. The availability reflects the accumulated evidence for how appropriate it would be for sample $\mathbf{p}_i$ to choose sample $\mathbf{p}_k$ as its exemplar, taking into account the support from other samples that sample $\mathbf{p}_k$ should be an exemplar. The availabilities are initialized to zero, and the pose responsibilities are then computed iteratively using the rule of Eq.2.3. The availabilities are updated in each iteration using Eq.2.4.

$$r_p(i,k) = s_p(i,k) - \max_{k'|k' \neq k} \{a_p(i,k') + s_p(i,k')\} \ , \tag{2.3}$$

$$a_p(i,k) = \min\Big\{0, r_p(k,k) + \sum_{i'|i' \notin \{i,k\}} \max\{0, r_p(i',k)\}\Big\} \ . \tag{2.4}$$

For $\mathbf{p}_i$, the value of $\mathbf{p}_k$ that maximizes $a_p(i,k) + r_p(i,k)$ either identifies sample $\mathbf{p}_i$ as an exemplar if $k = i$, or identifies the sample that is the exemplar for the sample $\mathbf{p}_i$. The message-passing procedure is terminated after a fixed number of iterations when the local cost functions remain constant for some number of iterations. At the end of pose clustering, K pose clusters $\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_j, \ldots, \mathbf{P}_K\}$ are determined, where $\mathbf{p}_i = [\theta_i^{pitch}, \theta_i^{yaw}, \theta_i^{roll}]$.

The second clustering is then applied for each pose cluster in the $l^{i,j}$ and $c^{i,j}$ measure space to find lighting exemplars. The first step computes illumination-contrast similarities $s_u(i,k) = - \parallel (\mathbf{u}^i - \mathbf{u}^k) \parallel^2$. The corresponding responsibility and availability are obtained according to:

$$r_u(i,k) = s_u(i,k) - \max_{k'|k' \neq k} \{a_u(i,k') + s_u(i,k')\} \ , \tag{2.5}$$

$$a_u(i,k) = \min\Big\{0, r_u(k,k) + \sum_{i'|i' \notin \{i,k\}} \max\{0, r_u(i',k)\}\Big\} \ . \tag{2.6}$$

The estimated $r_{cl}(i,k)$ and $a_{cl}(i,k)$ are combined to monitor the exemplar decisions and the algorithm is terminated when these decisions do not change for several iterations. At the end of the illumination-contrast clustering for each pose cluster $\mathbf{P}_j$, a number of N_j lighting clusters $\mathbf{P}_j = \{\mathbf{U}_{j1}, \mathbf{U}_{j2}, \ldots, \mathbf{U}_{jN_j}\}$ are obtained. The central representative samples of illumination-contrast clusters in j^{th} pose cluster are considered as the pose and lighting exemplars for j^{th} pose as $\mathbf{u}_j^i = (l_j^i, c_j^i), 1 \leq i \leq N_j$ where l and c are illumination and contrast of center of i^{th} illumination-contrast cluster $\mathbf{U}_{ji}$ in the j^{th} pose cluster $\mathbf{P}_j$.

Larger clusters represent a greater number of generic samples, they should have more influence for the classification. Therefore, a weight is assigned to each exemplar $\mathbf{u}_j^i$ to indicate its importance, approximated based on its cluster size, $W_{ij} = n_{ij}/n$, where n_{ij} is the number of samples in the cluster $\mathbf{U}_{ij}$ and n is the number of generic samples. This selection strengthens those classes that are more representative in reconstructing a probe sample.

2.3.2 Face Synthesis

For generating synthetic ROIs based on the representative pose and lighting conditions, 3D models of reference ROIs are reconstructed and their material-dependent layers are extracted. In the rendering process, the extracted material layers are employed as a texture of the 3D model. This model is rendered w.r.t. the pose exemplars. Following this, the illumination-dependent layers of the lighting exemplars are extracted. Finally, the lighting layers are projected on the rendered images with the same view by applying a morphing between the layers. The following subsections describe the steps proposed for the face synthesizing with DSFS.

2.3.2.1 Intrinsic Image Decomposition

Each still reference image, $\mathbf{r}_i$, is decomposed to its material-dependent layer (albedo), $\mathbf{M}_i$, and shading-dependant layer, $\mathbf{L}_i$, based on the a texture-aware image model defined by Jeon *et al.* (2014). This image decomposition method explicitly models a separate texture layer in addition to the shading layer and material layer in order to avoid ambiguity caused by textures. Explicitly modeling textures, shading layer and reflectance layer in the model depict only textureless base components, and accordingly avoid ambiguity caused by textures. Furthermore, for robustness against noise, the points are sparsely sampled for the surface normal constraint based on local variances of surface normal. This model is presented as follows:

$$\mathbf{r}_i(x,y) = \mathbf{B}^i(x,y).\mathbf{T}^i(x,y) = \mathbf{L}^i(x,y).\mathbf{M}_i(x,y).\mathbf{T}^i(x,y) \ , \tag{2.7}$$

where $\mathbf{B}(x,y) = \mathbf{L}(x,y).\mathbf{M}(x,y)$ is a base layer, and $\mathbf{L}(x,y)$, $\mathbf{M}(x,y)$ and $\mathbf{T}(x,y)$ are shading, material, and texture components at a pixel (x,y), respectively.

2.3.2.2 3D Face Reconstruction

3D face model of reference ROIs, $\mathbf{r}_i$, are reconstructed using the 3DMM technique (Blanz & Vetter (2003); Paysan *et al.* (2009)). In this study, a customized version of the 3DMM is employed in which the texture fitting of the original 3DMM is replaced with image mapping. By replacing the texture fitting in the original 3DMM with 2D image mapping, an efficient method is implemented for 3D face reconstruction from one frontal face image. Basically, the shape model is defined as a convex combination of shape vectors of a set of examples in which the shape vector ($\mathbf{S}$) is defined as Eq.2.8 (Blanz & Vetter (2003)). A principal components analysis is performed to estimate the statistics of the 3D shape of the faces.

$$\mathbf{S} = \bar{\mathbf{S}} + \sum_{k=1}^{m_S-1} \alpha_k.\tilde{\mathbf{S}}_k \ , \tag{2.8}$$

where, the 3D shape is represented by the probability distribution of faces around the averages of shape $\bar{\mathbf{S}}$ and the basis vectors $\tilde{\mathbf{S}}_j$, $1 \leq j \leq m_s$ in Eq.2.8 where m_s is the number of the basis vectors.

Each vector $\mathbf{S}$ stores the reconstructed 3D shape in terms of x, y, z-coordinates of all vertices $\varepsilon\{1,\ldots,n_s\}$ of a high-resolution 3D mesh as

$$\mathbf{S} = [X_1, Y_1, Z_1, X_2, \ldots, X_{n_s}, Y_{n_s}, Z_{n_s}]^T \ . \tag{2.9}$$

Here, for each reference ROI, $\mathbf{r}_i$, we reconstruct the 3D shape.

$$\mathbf{S}_i = \bar{\mathbf{S}} + \sum_{j=1}^{m_S-1} \alpha^i_j.\tilde{\mathbf{S}}_j, \tag{2.10}$$

where $\alpha^i_j \in [0,1], 1 \leq j \leq m_s$ are the shape parameters and $\mathbf{S}_i$ is the reconstructed shape of the i^{th} reference still ROI $\mathbf{r}_i$. The optimization algorithm presented by Blanz & Vetter (2003) is employed to find optimal $\alpha^i_j, 1 \leq j \leq m_s$, for each reference still ROI $\mathbf{r}_i$. In the next step, the extracted material layers, $\mathbf{M}_i$, are projected to the 3D geometry of the reference gallery set. Given the 3D facial shape and texture, novel poses can be rendered under various forms of the pose by adjusting the parameters of a camera model. In the rendering procedure, the 3D face is projected onto the image plane with Weak Perspective Projection which is a linear approximation of the full perspective projection.

Since the 2D image is directly mapped to the 3D model, no corresponding color information is available for some vertices because they are occluded in the frontal face image. Consequently, it is possible that there are still some blank areas on the generated texture map. In order to correct these blank space areas, a bilinear interpolation algorithm is utilized to fill in areas of unknown texture using the known colors in the vicinity.

2.3.2.3 Illumination Transferring

For each pose exemplar, $\mathbf{p}_j$, a set of samples $\{\mathbf{I}_{\mathbf{u}}^{jk} \in \mathbb{R}^{d\times d} | 1 \leq k \leq N_j\}$ corresponding to the $\mathbf{u}_{jk}$, for $k = 1, 2, ..., N_j$, are selected as lighting exemplars. The illumination-dependent layer of each $\mathbf{I}_{\mathbf{u}}^{jk}$ are extracted using the same process described in section 2.3.2.1. For each pose exemplar $\mathbf{p}_j$, N_j the illumination layers, $\mathbf{L}_{jk}$, for $k = 1, 2, ..., N_j$, are then projected to the rendered reference, $\mathbf{V}_{ij}$. This is performed by morphing between $\mathbf{L}_{jk}$ and $\mathbf{V}_{jk}$ according to the following steps:

i) detect the landmark points of $\mathbf{L}_{jk}$ and $\mathbf{V}_{ij}$ using active shape model to locate corresponding feature points. The landmark points of $\mathbf{L}_{jk}$ and $\mathbf{V}_{ij}$ are denoted as l_{jk} and v_{ij}, respectively;

ii) define a triangular mesh over l_{jk} and v_{ij} via the Delaunay triangulation technique and obtain d_l^{jk} and d_v^{ij};

iii) coordinate transformations between d_l^{jk} and d_v^{ij} with affine projections on the points;

iv) warp each triangle separately from the source to destination using mesh warping technique which moves triangular patches to the newly established location to align two ROIs;

v) cross-dissolve the triangulated layers considering warped pixel locations.

In this way, a number $q = \sum_{j=1}^{K} N_j$ of synthetic ROIs are generated for each reference still ROI $\mathbf{r}_i$. Therefore, the total number of synthetic ROIs are $q_{total} = nq$. The synthetic set of ROIs for the i^{th} reference still ROI are presented by $\mathbf{S}^i = [\mathbf{s}_1^i, \mathbf{s}_2^i, \ldots, \mathbf{s}_q^i] \in \mathrm{I\!R}^{\mathrm{d}^2 \times \mathrm{q}}$ where $\mathbf{s}_j^i$ is the j^{th} concatenated synthetic ROI for the i^{th} reference still ROI. The overall process of DSFS face generation technique is formalized in Algorithm 2.1.

Algorithm 2.1 The DSFS Approach.

```
Input: Reference set $\mathbf{R} = \{\mathbf{r}_i \in \mathbb{R}^{d \times d} | i = 1, \ldots, n\}$, and generic set
       $\mathbf{G} = \{\mathbf{g}_i \in \mathbb{R}^{d \times d} | i = 1, \ldots, m\}$.
Estimate pose angles.
Calculate luminance and contrast distortion measures. // Eq.2.1, Eq.2.2
AP clustering on pose space to obtain $\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_K\}$. // Section2.3.1.3
for $j = 1$ to $K$ do
    AP clustering on Illumination and contrast space for the $\mathbf{P}_j$ to obtain $\{\mathbf{u}_{ji} | 1 \leq i \leq N_j\}$.
     // Section2.3.1.3
end
for $i = 1$ to $n$ do
    Extract material-dependent layer of $\mathbf{r}_i$ ($\mathbf{M}_i$). // Section2.3.2.1
    Recover 3D face model of $\mathbf{r}_i$ using 3DMM ($\mathbf{S}_i$). // Section2.3.2.2
    Map the texture of $\mathbf{M}_i$ to $\mathbf{S}_i$.
    for $j = 1$ to $K$ do
        Render under $\mathbf{p}^j$ pose to obtain $\mathbf{V}_{ij}$.
        for each $\mathbf{u}_{ji}$, $i = 1$ to $N_j$ do
            Extract illumination-dependent layers ($\mathbf{L}_{jk}$). // Section2.3.2.1
            Morphing between $\mathbf{L}_{jk}$ and $\mathbf{V}_{ij}$ to obtain $\mathbf{S}^i$. // Section2.3.2.3
        end
    end
end
Output: All sets of synthetic face ROIs under representative pose and illumination conditions.
        $\mathbf{S}^i = [\mathbf{s}_1^i, \mathbf{s}_2^i, \ldots, \mathbf{s}_q^i] \in \mathrm{I\!R}^{\mathrm{d}^2 \times \mathrm{q}}$, $\mathrm{i} = 1, 2, \ldots, n$ .
```

2.4 Domain-invariant Face Recognition with DSFS

In this section, a particular still-to-video FR implementation is considered (see Fig. 2.5) to assess the impact of using DSFS to generate synthetic ROIs to address these limitations.

An augmented dictionary is constructed by employing the synthetic ROIs generated via DSFS technique, and classification is performed via a structured SRC approach. Since the synthetic

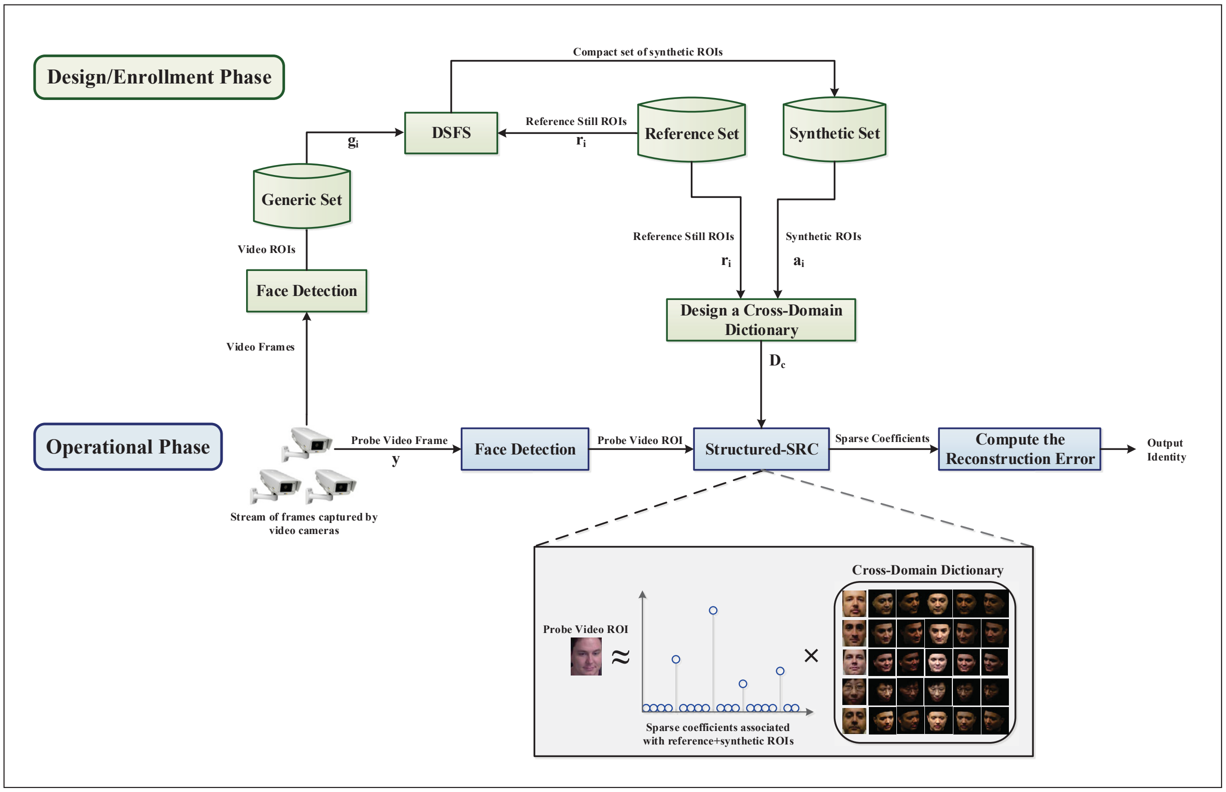


Figure 2.5 Block diagram of the proposed domain-invariant SRC-based still-to-video FR system.

ROIs for each individual (including the synthetic poses, illuminations, and etc.) form a block in this dictionary, the SRC is considered as a structured sparse recovery problem. The main steps of the proposed domain-invariant still-to-video FR with dictionary augmentation are summarized as follows:

- **Step 1**: Generation of Synthetic Facial ROIs

  In the first step, q synthetic ROIs $\mathbf{S}^i = [\mathbf{s}^i_1, \ldots, \mathbf{s}^i_q] \in R^{d^2 \times q}$ are generated for each $\mathbf{r}_i$ of the reference gallery set using DSFS technique, where q is the number of synthetic ROIs for each class.

- **Step 2**: Augmentation of Dictionary

  The synthetic ROIs generated through the DSFS technique are added to the reference dictionary to design a cross-domain dictionary. Let $\mathbf{D}_R = [\mathbf{I_r}^1, \mathbf{I_r}^2, \ldots, \mathbf{I_r}^n] \in \mathbb{R}^{d^2 \times n}$ be the reference gallery dictionary, where $\mathbf{I_r}^i$ is the concatenated result of $\mathbf{r}_i$. The cross-domain dictionary $\mathbf{D}_C = [\mathbf{I_r}^1, \mathbf{S}^1, \ldots, \mathbf{I_r}^n, \mathbf{S}^n] \in \mathbb{R}^{d^2 \times n(q+1)}$ integrates the original and synthetic ROIs in a linear model where $\mathbf{S}^j$ is the j^{th} set of synthetic ROIs added to the j^{th} class. Since q synthetic ROIs are added to each class, the total number of ROIs in the cross-domain dictionary are $n_c = n(q+1)$.

The presented dictionary design in this work enables SRC to perform recognition with only one reference still ROI and makes it robust to the visual domain shift.

- **Step 3**: Classification

  Given a probe video ROI $\mathbf{y}$, general SRC represents $\mathbf{y}$ as a sparse linear combination of the codebook $\mathbf{D}_C$, and derives the sparse coefficients of $\mathbf{y}$ by solving the ℓ_0-minimization problem as follows:

$$A_{\ell_0}: \quad \min \|\mathbf{x}\|_0 \quad s.t. \quad \mathbf{y} = \mathbf{D}_C \mathbf{x}\ . \tag{2.11}$$

  Since the generated synthetic ROIs for each individual form a block of the dictionary, a better classification can arise from a representation of the probe ROI produced from the minimum number of blocks from the dictionary instead of looking for the representation of a probe ROI in the dictionary of all the training data using the so-called structured SRC which its goal is to find a representation of a probe ROI that uses the minimum number of blocks from the dictionary. For a dictionary $\mathbf{D}_C = \left[\mathbf{D}_C[1], \mathbf{D}_C[2], \ldots, \mathbf{D}_C[n]\right]$ with blocks $\mathbf{D}_C[i]$, $i = 1, \ldots, n$, the block sparsity is formulated in terms of mixed ℓ_2/ℓ_0 norm as;

$$A_{\ell_2/\ell_0}: \quad \min_{\mathbf{x}} \sum_{i=1}^{n} I(\parallel \mathbf{x}[i] \parallel_2 > 0) \quad s.t. \quad \mathbf{y} = \mathbf{D}_C \mathbf{x}\ , \tag{2.12}$$

  where $I(.)$ is the indicator function, and $\mathbf{x}[i]$ is the i^{th} block in the sparse coefficient vector $\mathbf{x}$ corresponding to the dictionary block $\mathbf{D}_C[i]$. Since each dictionary block corresponds to a specific class, i represents the class index ranging from 1 to n as well. This optimization problem seeks the minimum number of non-zero coefficient blocks that reconstruct the probe ROI.

Note that the optimization program A_{ℓ_2/ℓ_0} is NP-hard since it requires searching over all possible few blocks of **x** and checking whether they span the given **y**. A relaxation of this problem is obtained by replacing the ℓ_0 with the ℓ_1 norm and solving the Eq.2.13.

$$A_{\ell_2/\ell_1}: \quad \min_{\mathbf{x}} \sum_{i=1}^{n} \| \mathbf{x}[i] \|_2 \quad s.t. \quad \mathbf{y} = \mathbf{D}_C \mathbf{x} \,. \tag{2.13}$$

Finally, the weighted matrix obtained in 2.3.1.3 which shows cluster weights is multiplied to the ℓ_1-minimization term.

$$A_{\ell_2/\ell_1}: \quad \hat{x} = arg \min_{\mathbf{x}} \sum_{i=1}^{n} \|\mathbf{W}_i \, \mathbf{x}[i] \, \|_2 \quad s.t. \quad \mathbf{y} = \mathbf{D}_C \mathbf{x} \,. \tag{2.14}$$

where

$$\mathbf{W}_i = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{i(q+1)} \end{bmatrix} . \tag{2.15}$$

In order to solve the SRC problem of equation 2.14, the classical alternating direction method (ADMM) is considered which is an efficient first-order algorithm with global convergence (Deng *et al.* (2013)).

The class label of the probe ROIs y is then determined based on the reconstruction error as follows:

$$\text{label}(\mathbf{y}) = arg \min_{i} \| \, \mathbf{y} - \mathbf{D}[i]\hat{\mathbf{x}}[i] \, \|_2 \,. \tag{2.16}$$

- **Step 4**: Validation

  In practical FR systems, it is important to detect and then reject outlier invalid probe ROIs. We use the *sparsity concentration index (SCI)* criteria defined by Wright *et al.* (2009):

$$\text{SCI}(\hat{\mathbf{x}}) \doteq \frac{n.\max_i \| \, \hat{\mathbf{x}}[i] \, \|_1 \, / \, \| \, \hat{\mathbf{x}} \, \|_1 \, -1}{n-1} \quad \in [0,1] \,. \tag{2.17}$$

where n is the number of classes. A probe ROI is accepted as valid if $SCI \geq \tau$ and otherwise rejected as invalid, where $\tau \in (0,1)$ is a threshold.

The still-to-video FR process through dictionary augmentation is formalized in Algorithm 2.2.

Algorithm 2.2 A SRC-based Still-to-Video FR System.

```
Input: Reference face models of $n$ classes enlisted in the gallery $\mathbf{R} = \{\mathbf{r}_i \in \mathbb{R}^{d \times d} | i = 1, \ldots, n\}$,
       generic set $\mathbf{G} = \{\mathbf{g}_i \in \mathbb{R}^{d \times d} | i = 1, \ldots, m\}$, threshold $\tau$, and a probe ROI $\mathbf{y}$.
Generate $nq$ synthetic ROIs for each class using the DSFS method.
Build the cross-domain dictionary $\mathbf{D}_C$ by adding the synthetic ROIs to the reference gallery set.
Solve the $A_{\ell_2/\ell_1}$ problem using ADM technique. if $SCI \geq \tau$ then
                                                                    // Eq.  2.17
    Find the class label $\mathbf{y}$.                              // Eq.  2.16
else
    Reject as invalid.
end
Output: Class label of $\mathbf{y}$.
```

2.5 Experimental Methodology

2.5.1 Databases

In order to validate the proposed DSFS for still-to-video FR under real-world surveillance conditions, extensive experiments were conducted on two publicly available datasets – COX-S2V (Huang *et al.* (2015)) and Chokepoint (Wong *et al.* (2011)). These datasets were selected because they are the most representative for watch-list screening applications. They contain a high-quality reference image per subject captured under controlled condition (with a still camera), and lower-quality surveillance videos for each subject captured under uncontrolled conditions (with surveillance cameras).

COX-S2V dataset contains 1000 individuals (435 male and 565 female), with 1 high-quality still image and 4 low-resolution video sequences per individual simulating video surveillance scenario. In each video, an individual walk through a designed S-shape route with changes in illumination, scale, and pose (Huang *et al.* (2015)). The Chokepoint dataset consists of 25 individuals (19 male and 6 female) walking trough portal 1, and 29 individuals walking trough portal 2. The recording of portal 1 and portal 2 are one month apart. A camera rig with 3 cameras is used for simultaneously recording the entry of a person during four sessions with changes in illumination conditions, pose, and misalignment. In total, the dataset consists of 54 video sequences and $64,204$ face images (Wong *et al.* (2011)).

2.5.2 Experimental protocol

With the Chokepoint database, 5 individuals are randomly chosen as watch-list individuals that each individual includes a high-quality frontal captured image. Prior to each experiment, the video data is split into 3 parts. ROIs are extracted from the video sequences of 10 other individuals selected at random as a generic set to represent capture conditions. ROIs of the video sequences of the remaining individuals along with video sequences of the 5 already selected watch-list individuals are employed for testing. In order to obtain representative results, this process is repeated 5 times with a different random selection of watch-list and generic set individuals, and the average accuracy is reported with mean and standard deviation over all the runs. With COX-S2V, 30 individuals are randomly considered as watch-list individuals including a high-quality captured image per each individual. Their corresponding low-quality video sequences along with ROIs of the video sequences of 100 other individuals are employed for testing. The ROIs extracted from the video sequences of 100 other individuals are selected at random as a generic set to represent capture conditions. This process is replicated 5 times with different stills and videos of watch-list individuals, and the average accuracy is reported with mean and standard deviation over all the runs. During the enrollment, the ROIs of the generic set of faces captured from video trajectories across all target domains (i.e., global modeling) are extracted using the Viola-Jones face detection algorithm (Viola & Jones (2004)). Face detection is also applied to still images prior to face synthesis. An estimation of luminance and contrast are measured for each video ROIs in the generic set where the constant values of K_l and K_C are set to 0.01 and 0.03, respectively as proposed by Wang *et al.* (2004). Pose angles are estimated. Then, AP clustering is applied to the generic set, where q representative video ROIs are selected under various pose, illumination and contrast conditions, and a weight is assigned to each exemplar according to the cluster size. Then, q synthetic face images are generated for each individual based on the information obtained from these selected exemplars. Recall that AP clustering seeks exemplars (samples that are representative of clusters), and automatically determines k and q, the number of clusters, for each independent replication. The cross-domain dictionary is then designed using the reference still and synthetic ROIs. During the operational phase, recognition is performed by coding the probe image over the cross-domain dictionary regarding the weights obtained in the source domain. Throughout the experiments, the sparsity parameter λ is fixed to 0.005. For reference, the still-to-video FR system based on individual-specific SVMs is also evaluated. During the enrollment, a non-linear SVM classifier with RBF kernel is trained for each individual using target ROIs (reference still of the individual plus the related synthetically face images) versus non-target ROIs (reference still of cohort persons plus their synthetic face images).

2.5.3 Performance Measures

To assess the ability of face synthesizing techniques to address shifts between target domain and source domain, a domain shift quantification (DSQ) measure is employed. With this measure, the similarity between a dictionary designed using synthetic ROIs ($\mathbf{D}_A$) is compared with a dictionary formed with images collected from the target domain ($\mathbf{D}_R$) by measuring the mean pixel error between the dictionaries. Given two dictionaries $\mathbf{D}_A$ and $\mathbf{D}_R$ with the same number of images, the DSQ measure is defined as $Q_{dsq} = \|\mathbf{D}_R^T \mathbf{D}_A\|_F$ where a higher value indicates less domain shift (Ni *et al.* (2013)). The accuracy of the still-to-video FR system is assessed per individual of interest at the transaction level, using the receiver operating characteristic (ROC) space, where the true positive rates (TPRs) are plotted as a function of false positive rates (FPRs) over all threshold values. TPR is the proportion of target ROIs that correctly classified as individuals of interest over the number of target ROIs, while FPR is the proportion of non-target ROIs incorrectly classified as individuals of interest over the number of non-target ROIs. The area under ROC curve is a global scalar measure of accuracy that can be interpreted as the probability of correct classification over the range of TPR and FPR. Accordingly, accuracy of FR systems is estimated using the partial area under ROC curve pAUC(10%) (using the AUC at $0 <$FPR$\leq 0.1\%$). Since the number of target and non-target data are imbalanced, the area under precision-recall curves (AUPR) is also used to estimate the performance of FR systems.

2.6 Results and Discussion

This section first presents some examples of synthetic faces generated using the DSFS technique and compares them with synthetic faces generated using state-of-the-art face synthesizing methods: 3DMM (Blanz & Vetter (2003)), and 3DMM-CNN (Tran *et al.* (2017a)). Then, the performance of still-to-video FR systems based on individual-specific SVMs and on SRC is presented when using these synthetic facial ROIs for system design. FR performance is assessed when increasing the number of synthetic ROIs per each individual according to pose angles and lighting effects. To characterize the impact on performance, these systems are tested with a growing number of synthetic ROIs and generic training set, and compared with several relevant state-of-the-art still-to-video FR systems: ESRC (Deng *et al.* (2012)), RADL (Wei & Wang (2015)), SVDL (Yang *et al.* (2013)), LGR (Zhu *et al.* (2014)), and Flow-based face frontalization (Hassner *et al.* (2015)). The final experiment compares the performance of a system designed with synthetic ROIs obtained with DSFS, to a system designed with a growing number of randomly-selected synthetic ROIs. The dataset,face synthesizing and face recognition experiments can be viewed at https://github.com/faniamokhayeri/DSFS.

2.6.1 Face Synthesis

This subsection presents examples of pose and lighting exemplars obtained by clustering of facial trajectories in the captured condition space. Fig. 2.6 shows an example of pose clusters obtained with Chokepoint video trajectories of 10 individuals, and with COX-S2V video trajectories of 100 individuals. In this experiment, $k_1 = 9$ and $k_2 = 7$ pose clusters (exemplars) are typically determined with the Chokepoint and COX-S2V videos, respectively. The second level of clustering is then applied in the illumination and contrast measure space on each pose clusters. Fig. 2.7 shows the exemplars selected based on both pose and lighting with the proposed representative selection of DSFS (see section 2.3.1). Overall, $q_1 = 22$ and $q_2 = 18$ exemplars were typically selected based on both pose and lighting clusters determined in Chokepoint and COX-S2V videos, respectively.

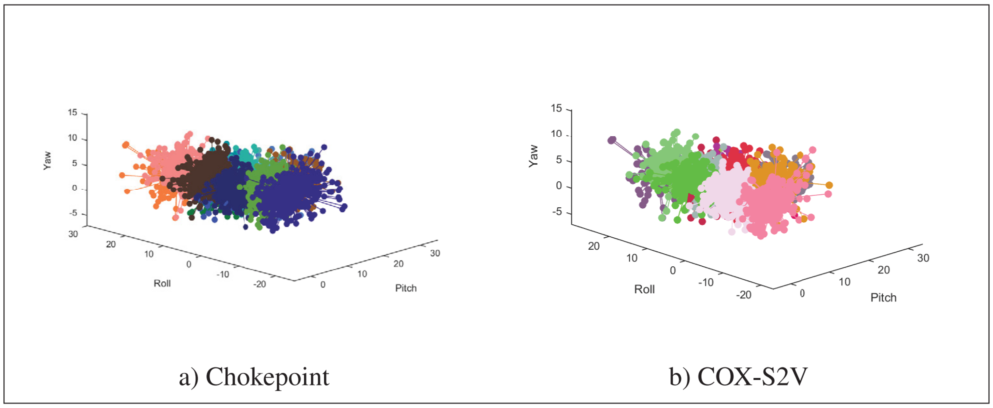


Figure 2.6 Examples of representative selection results using AP clustering technique in terms of pose angles with Chokepoint dataset on video sequences of 10 individuals and COX-S2V dataset on video sequences of 100 individuals, respectively.

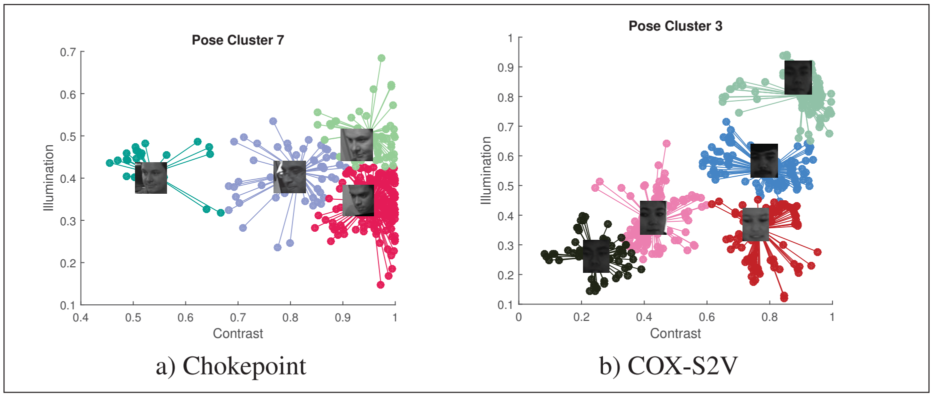
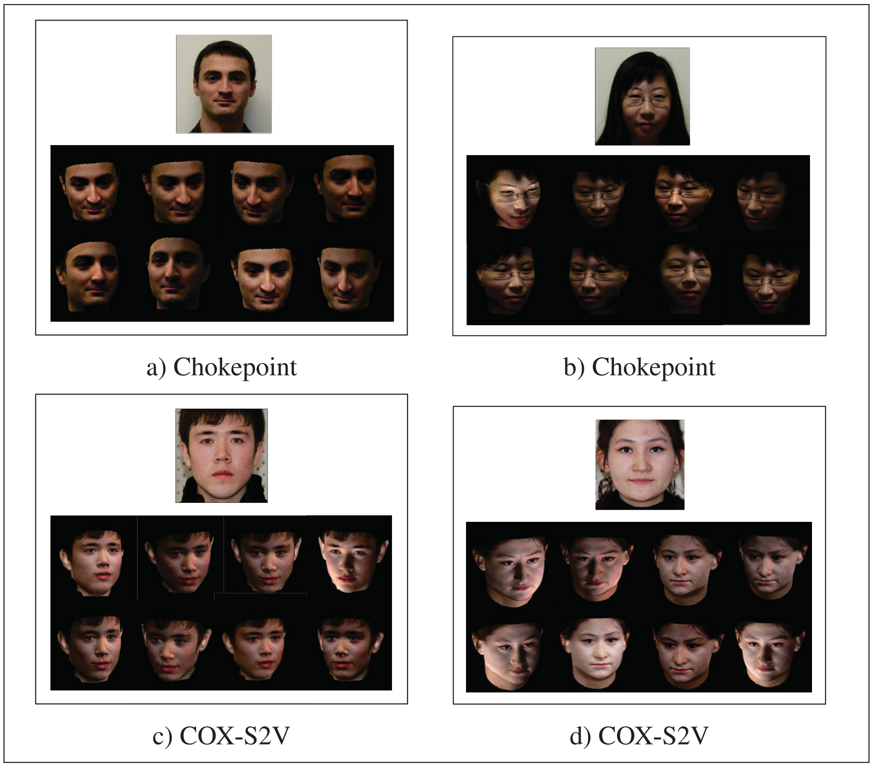


Figure 2.7 Examples of luminance and contrast representatives with Chokepoint dataset on video sequences of 10 individuals and COX-S2V dataset on video sequences of 100 individuals, respectively, where the center of clusters show exemplars.

Fig. 2.8 show examples of synthetic ROIs generated under different pose, illumination and contrast conditions using the DSFS technique on the Chokepoint and COX-S2V datasets, where

Basel Face Model are used as generative 3D shape model (Paysan *et al.* (2009)). In Fig.2.9, the quality of synthetic faces generated under different pose via DSFS, 3DMM (Blanz & Vetter (2003)), and 3DMM-CNN (Tran *et al.* (2017a)) techniques are compared.



a) Chokepoint

b) Chokepoint

c) COX-S2V

d) COX-S2V

Figure 2.8 Examples of synthetic ROIs generated under different capture conditions using the DSFS technique with Chokepoint (a,b) and COX-S2V (c,d) datasets.

The synthetic ROIs generated using the DSFS are also evaluated quantitatively. Table. 2.1 shows the DSQ values of the DSFS and other state-of-the-art face synthesizing methods including 3DMM (Blanz & Vetter (2003)), 3DMM-CNN (Tran *et al.* (2017a)), and SHBMM (Zhang & Samaras (2006)) on Chokepoint and COX-S2V datasets. Higher DSQ values indicate a smaller domain shift, and potentially higher recognition rate between the corresponding two domains. The results are provided under the two following scenarios.

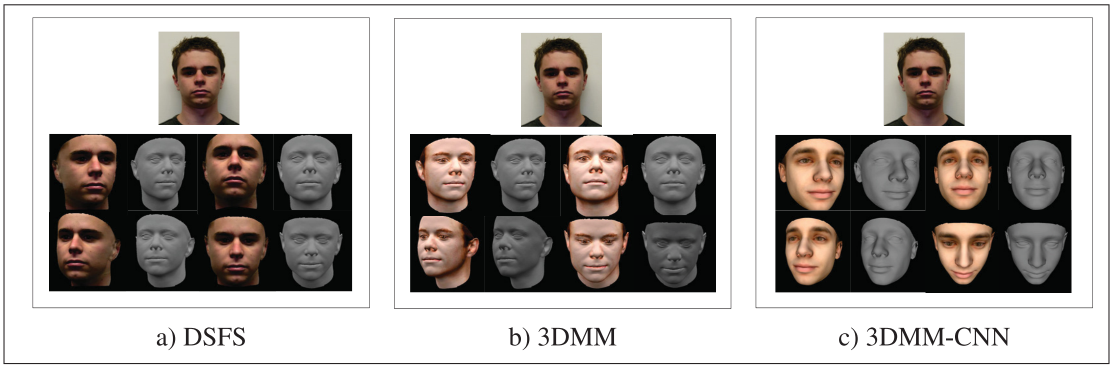


Figure 2.9 Synthetic face images generated under different pose via (a) DSFS, (b) 3DMM, (c) 3DMM-CNN with Chockpoint dataset.

2.6.1.1 Frontal View

In the first experiment, 5 individuals in source domain are randomly selected. A set of synthetic face are generated with a frontal view under various lighting effects from the still ROI of each individual to design $\mathbf{D}_A$. The corresponding video ROIs in the target domain under the frontal view are then collected to form $\mathbf{D}_R$. Finally, the DSQ is measured for the $\mathbf{D}_A$ and $\mathbf{D}_r$ dictionaries.

2.6.1.2 Profile View

In the second experiment, 5 individuals in source domain are again randomly selected. Their synthetic ROIs are generated with profile view and different illumination conditions to form $\mathbf{D}_A$. The corresponding video ROIs in target domain under profile view are collected to construct $\mathbf{D}_R$. Finally, the DSQ is estimated for the dictionaries.

As shown in Table. 2.1, DSQ values of DSFS method are higher followed most closely by SHBMM in both scenarios. Accordingly, the cross-domain dictionary designed by the synthetic ROIs generated via the DSFS method is most suitable to reduce visual domain shifts and potentially achieve a higher level of accuracy. These results are in line with the recognition performance results.

Table 2.1 Average DSQ value for frontal and profile views on Chokepoint and COX-S2V datasets.

Technique	**DSQ**			
	Chokepoint database		**COX-S2V database**	
	Frontal View	Profile View	Frontal View	Profile View
3DMM	8.27	7.38	8.24	7.63
3DMM-CNN	7.61	7.03	7.57	7.29
SHBMM	9.16	7.26	9.34	7.71
Proposed DSFS	**9.53**	**8.17**	**9.58**	**8.39**

2.6.2 Face Recognition

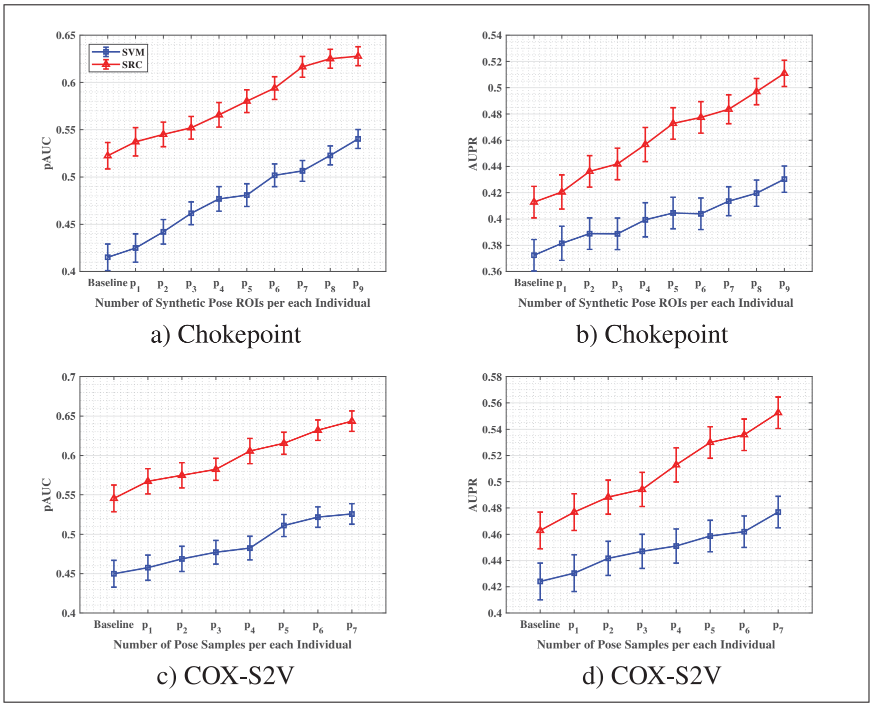
In this subsection, the performance achieved using the still-to-video FR system based on SRC and DSFS (see Section 2.3) is assessed experimentally. For reference, the still-to-video FR system based on individual-specific SVMs is also evaluated.

2.6.2.1 Pose Variations

The still-to-video FR system is evaluated versus the number of synthetic ROIs that incorporate growing facial pose. 2.10 show the average AUC and AUPR obtained by increasing the number of synthetic ROIs generated using DSFS from k representative pose angles ($\mathbf{p}_i$, $i = 1 \ldots k$) and with fixed lighting condition. Results indicate that by adding extra synthetic ROIs generated under representative pose angles allows to outperform baseline systems designed with an original reference still ROI alone. AUC and AUPR accuracy increases by about 10%, typically with only $k_1 = 9$ and $k_2 = 7$ synthetic pose ROIs for Chokepoint and COX-S2V datasets, respectively.

2.6.2.2 Mixed Pose and Illumination Variations

The performance of still-to-video FR systems is assessed versus the number of synthetic ROIs generated under both pose and lighting effects. Fig. 2.11 show average AUC and AUPR obtained by increasing the number of synthetic ROIs used to design SRC and SVM classifiers on the Chokepoint and COX-S2V databases, where $\mathbf{S}_i$ is a set of synthetic ROIs generated us-



a) Chokepoint

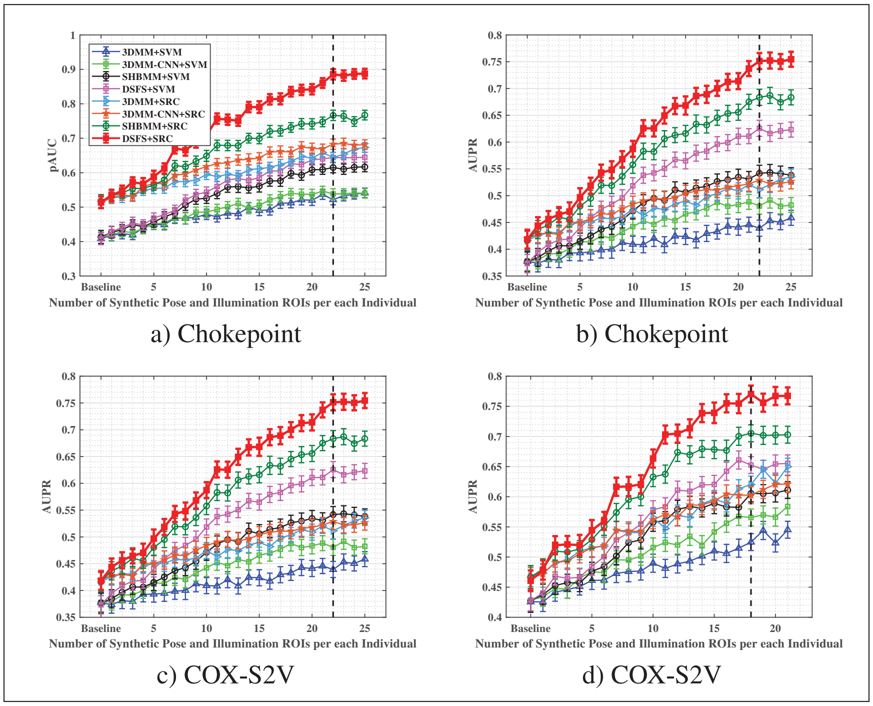
b) Chokepoint

c) COX-S2V

d) COX-S2V

Figure 2.10 Average AUC and AUPR versus the number of synthetic ROIs generated with DSFS according to various pose and fixed illumination. The still-to-video FR system employs either SVM and SRC classifiers on Chokepoint (a,b) and COX-S2V (c,d) databases.

ing DSFS technique under various pose and illumination conditions. Adding synthetic ROIs generated under various pose, illumination and contrast conditions allows to significantly outperform the baseline system designed with the original reference still ROI alone. AUC and AUPR accuracy increases by about 40%, typically with only $q_1 = 24$ and $q_2 = 18$ synthetic ROIs for Chokepoint and COX-S2V datasets, respectively. As shown in Fig. 2.11, accuracy for DSFS+SRC trends to stabilize to its maximum value when the size of the generic set is greater than q in DSFS. To view performance stabilizing with more than q synthetic ROIs, additional samples were selected randomly among AP clusters. Note that the Chokepoint dataset contains faces captured for a range illumination conditions with various densities. Hence, some exemplars may represent many video ROIs. Our method assigns higher weights to such distributions and may yield a higher level of performance. The results obtained with DSFS are also compared to still-to-video FR systems that exploit state-of-the-art face synthesis techniques including 3DMM (Blanz & Vetter (2003)) (with randomly selected images), and SHBMM (Zhang & Samaras (2006)) (with 9 spherical harmonic basis images). As shown in Fig. 2.11, DSFS always outperforms these other techniques.



a) Chokepoint

b) Chokepoint

c) COX-S2V

d) COX-S2V

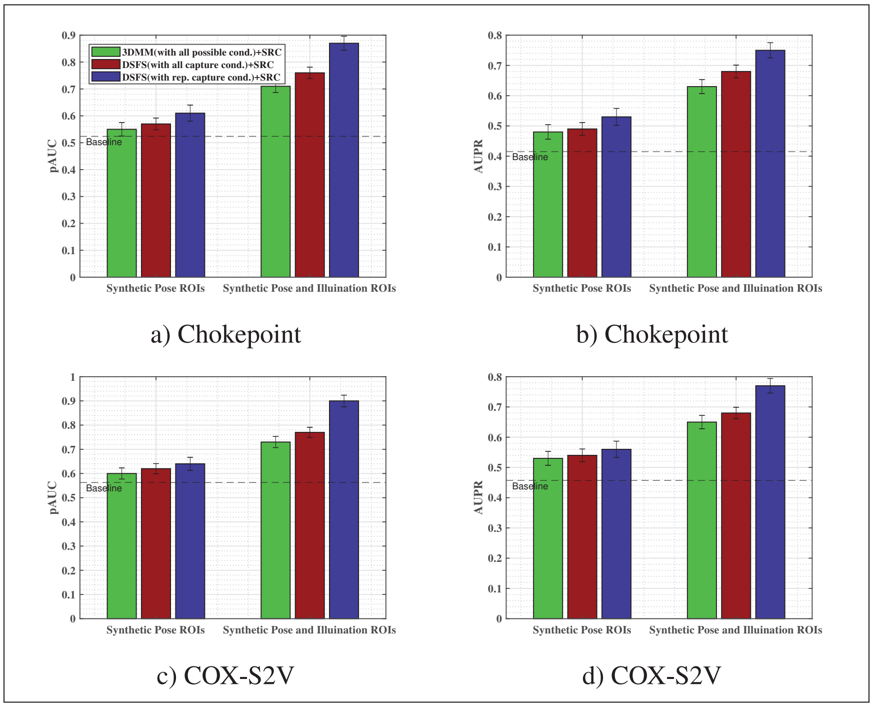
Figure 2.11 Average AUC and AUPR versus the number of synthetic ROIs generated with DSFS, 3DMM, and SHBMM according to pose and lighting effects where still-to-video FR system employs either SVM and SRC classifiers on Chokepoint (a,b) and COX-S2V (c,d) databases.

2.6.2.3 Impact of Representative Selection

Without prior knowledge of the target domain, synthetic faces are generated according to a uniform distribution. Adding a large number of synthetic ROIs to the dictionary as needed to cover all possible cases can significantly increase the time and memory complexity of FR systems and, more importantly, may cause over-fitting. The proposed DSFS technique extracts representative information from the target domain to produce a compact set of synthetic ROIs that are robust to intra-class variations in the target domain. In order to evaluate the impact of the synthetic ROIs generated based on representative information (i.e., pose and lighting cluster instances), 3 dictionaries are designed for SRC: (1) a dictionary designed with representative synthetic ROIs (DSFS technique); (2) a dictionary designed with the synthetic ROIs under all capture conditions (DSFS without AP clustering); and (3) a dictionary designed under all possible conditions.

The first scenario evaluates the impact of representative selection in terms of *pose* with 3 dictionaries. The first dictionary typically employs $k_1 = 9$ and $k_2 = 7$ representative synthetic pose ROIs generated with the DSFS technique for Chokepoint and COX-S2V datasets, respectively. The second dictionary employs 100 synthetic pose ROIs generated by DSFS technique under all target domain capture conditions. The third dictionary employs 180 synthetic pose ROIs generated by 3DMM in a set of rotation angles ranging from to -60 to $+60$ (2.12). The second scenario evaluates the impact of representative selection in terms of both *pose and illumination conditions* with 3 dictionaries designed for SRC. The first dictionary employs $q_1 = 22$ and $q_2 = 18$ representative synthetic ROIs generated under different pose and illumination with the DSFS technique for Chokepoint and COX-S2V datasets, respectively. The second dictionary employs 100 synthetic ROIs generated under different pose and illumination by DSFS technique under all target domain capture conditions. The third dictionary employs 180 synthetic pose ROIs generated under different pose and illumination by 3DMM (2.12). The results in Fig. 2.12 suggest that augmenting the dictionary using representative synthetic ROIs with the DSFS technique yields a higher level of accuracy, particularly under both pose and illumination conditions.

The impact of the proposed representative selection technique is also assessed based on the various pose estimation methods including DRMF (Asthana *et al.* (2013)), ERT (Kazemi & Josephine (2014)), and OpenFace (Baltruvsaitis *et al.* (2016)). For this, the performance of the FR system is compared according the different pose estimation techniques under combined variations of identity, pose, and illumination conditions. In this experiment, the 5 individuals of Chokepoint database and 30 individuals of COX-S2V database are used. With Chokpoint dataset, $q_1 = 22$, $q_2 = 20$, $q_3 = 24$ representative samples with DRMF, ERT, and OpenFace are obtained, re-



a) Chokepoint

b) Chokepoint

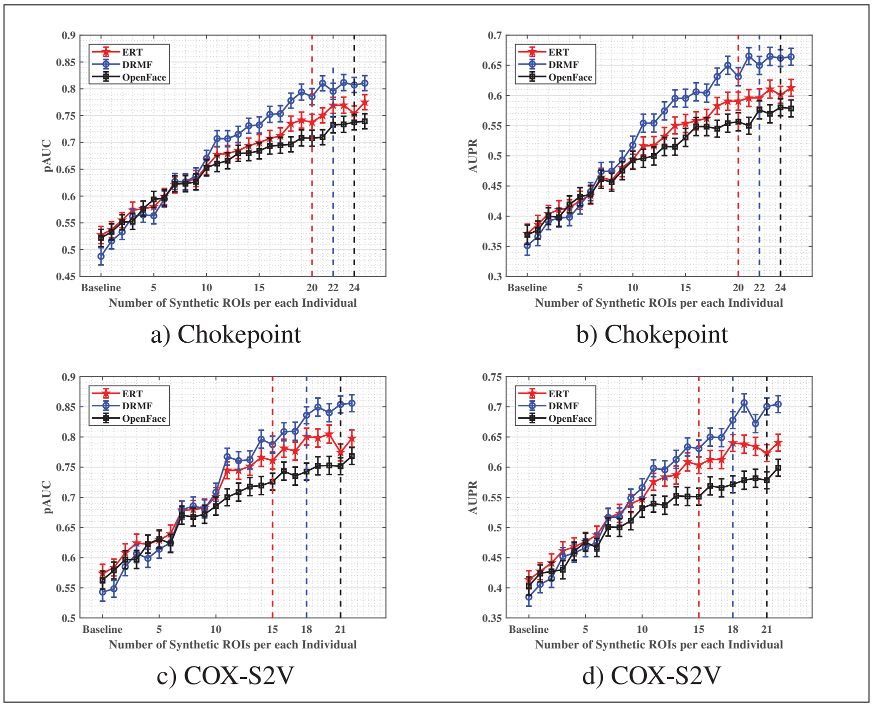
c) COX-S2V

d) COX-S2V

Figure 2.12 Average AUC and AUPR of a still-to-video FR system designed with representative synthetic ROIs vs a system designed with randomly generated synthetic ROIs on Chokepoint (a,b) and COX-S2V (c,d) datasets.

spectively, and With COX-S2V dataset, $q_1 = 18$, $q_2 = 16$, $q_3 = 21$ representative samples with DRMF, ERT, and OpenFace are obtained, respectively. Fig.2.13 show the average AUC and AUPR obtained by increasing the number of synthetic ROIs generated from q_1, q_2, and q_3 representative pose angles obtained with different pose estimation techniques. Then, the error of each pose estimation technique is computed based on the normalized distance of each landmark to its ground truth position (see Table. 2.2). It can be observed from the results that when error of pose estimation is low, the accuracy increases. The results suggest that the robustness of pose estimation techniques to nuisance factors has an impact on the performance of the FR system. The results are also compared with the situation where there is no pose estimation and the pose angles for face synthesizing are selected randomly.

The impact of illumination transferring on the DSFS technique is further evaluated. For this, the DSQ value (2.5.3) of the DSFS technique is compared based on the proposed illumination



a) Chokepoint

b) Chokepoint

c) COX-S2V

d) COX-S2V

Figure 2.13 Average AUC and AUPR accuracy obtained by increasing the number of the synthetic ROIs generated using DSFS from different representative pose angles (obtained with different pose estimation techniques) on Chokepoint (a,b) and COX-S2V (c,d) datasets.

Table 2.2 Average error rate of pose estimation for frontal and profile views on Chokepoint and COX-S2V datasets.

Technique	**Error rate of pose estimation**			
	Chokepoint database		**COX-S2V database**	
	Frontal View	Profile View	Frontal View	Profile View
DRMF	0.35	0.46	0.31	0.41
ERT	0.39	0.52	0.35	0.46
OpenFace	0.42	0.55	0.38	0.48

transferring method and the method presented by Chen *et al.* (2013) that transfer illumination through adaptive layer decomposition. In this experiment, we consider 5 and 30 individuals of Chokepoint and COX-S2V databases, respectively. With the proposed technique, DSQ of the Chokpoint and COX dataset are $DSQ = 8.63$ and $DSQ = 8.81$, respectively. With the adaptive layer decomposition method, DSQ of the Chokpoint and COX dataset are $DSQ = 7.24$ and

$DSQ = 7.39$, respectively. It can be concluded that the robustness of illumination transferring to unrelated distortions has an impact on the performance of DSFS technique. Since the shading decomposition technique employed in our illumination transferring technique is able to explicitly model the texture layer, the decomposed shading layer does not have any textures (Jeon *et al.* (2014)). As a result, It can avoid ambiguity caused by textures. However, weighted least squares filter employed by Chen *et al.* (2013) cannot deal with nuisance factors.

2.6.3 Comparison with Reference Techniques

With the above experimental setting, we compare the recognition rate of the DSFS technique with 3DMM (Blanz & Vetter (2003)) and SHBMM (Zhang & Samaras (2006)) methos in a still-to-video FR framework. We also present the impact of using face synthesizing along with KSVD dictionary learning (Aharon *et al.* (2006)).

Following this, the recognition rate of the DSFS technique with existing generic learning techniques including ESRC (Deng *et al.* (2012)), RADL (Wei & Wang (2015)), SVDL (Yang *et al.* (2013)), LGR (Zhu *et al.* (2014)) is compared that regularization parameter λ is set to 0.005. Note that the performance of the face synthesizing techniques is evaluated w/o dictionary learning. We also compared the DSFS results with the results obtained by Flow-based face frontalization method (Hassner *et al.* (2015)). Table. 2.3 lists and compares the recognition performance where the results (recognition rate) are illustrated by the mean and standard deviation of 5 runs.

2.6.3.1 Generic Set Dimension

In this subsection, the results of DSFS technique and some generic learning techniques are evaluated based on the size of the generic set. Given N generic images in the target domain, the recognition rate of the approaches is compared with increasing value of N. In this comparison, each system is considered as a black box, and their recognition rate is shown for a range of different numbers of inputs. 2.14 shows that for many generic learning techniques, intra-class

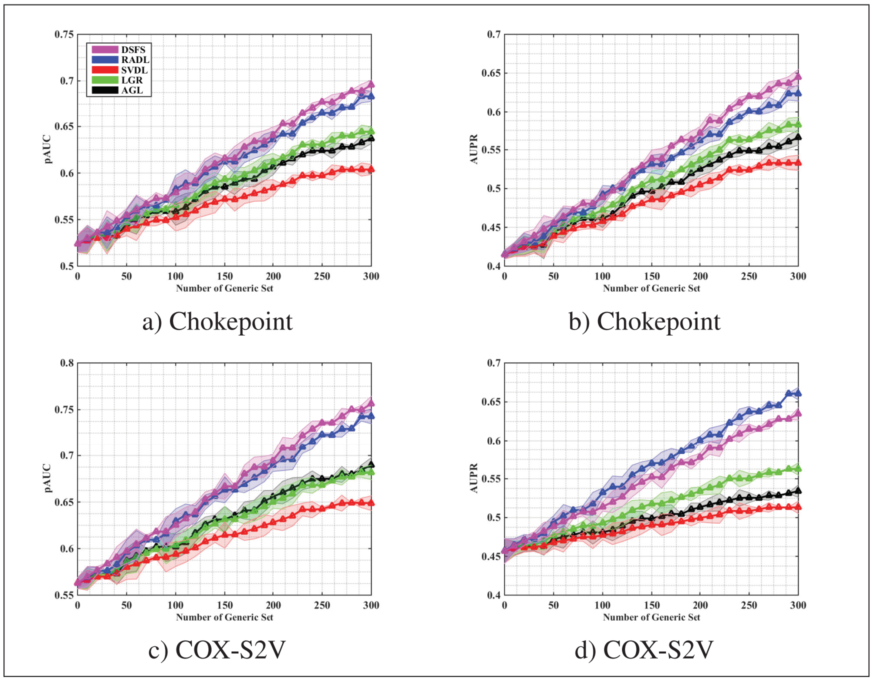
Table 2.3 Comparative transaction level analysis of the proposed FR approach and related state-of-the art FR methods with Chokepoint and COX-S2V databases.

Category	Technique	Classifier	Chokepoint database		COX-S2V database	
			pAUC	AUPR	pAUC	AUPR
	Baseline	SRC	0.516±0.033	0.415±0.035	0.548±0.031	0.457±0.036
Generic Learning	ESRC	SRC	0.798±0.029	0.651±0.032	0.827±0.028	0.695±0.032
		SRC-KSVD	0.809±0.024	0.672±0.022	0.831±0.018	0.715±0.020
	RADL	SRC	0.847±0.025	0.724±0.031	0.883±0.024	0.753±0.027
	LGR	SRC	0.841±0.028	0.717±0.024	0.877±0.026	0.744±0.025
	SVDL	SRC	0.823±0.021	0.703±0.029	0.839±0.022	0.724±0.031
Face Synthesizing	3DMM	SRC	0.663±0.035	0.523±0.037	0.702±0.031	0.562±0.032
		SRC-KSVD	0.712±0.032	0.605±0.032	0.732±0.031	0.641±0.028
	3DMM-CNN	SRC	0.672±0.025	0.516±0.026	0.705±0.025	0.552±0.025
		SRC-KSVD	0.716±0.024	0.585±0.025	0.741±0.026	0.603±0.025
	SHBMM	SRC	0.721±0.032	0.593±0.040	0.735±0.032	0.607±0.041
		SRC-KSVD	0.773±0.026	0.671±0.022	0.784±0.027	0.681±0.028
Face Frontalization	Flow-based	SRC	0.822±0.021	0.711±0.024	0.843±0.022	0.719±0.023
Face Synthesizing **+** **Generic Learning**	Proposed DSFS	SRC	**0.897±0.023**	**0.751±0.027**	**0.916±0.18**	**0.775±0.25**

variation of a small number of individuals in operational environment is sufficient to largely improve the recognition rate. In particular, it can be observed from Fig.2.14 that when more generic images are used, the accuracy increases significantly from our method and RADL technique (Wei & Wang (2015)), while the accuracies of other state-of-the-art methods do not change significantly. This shows that the proposed representative selection method is able to adequately select the representative faces out of a large set of faces.

Next, we compare the computational complexity in terms of average running time for each individual as well as number of inner products needed per each iteration. 2.15 shows the computational complexity in terms of number of inner products with a growing number of synthetic ROIs.

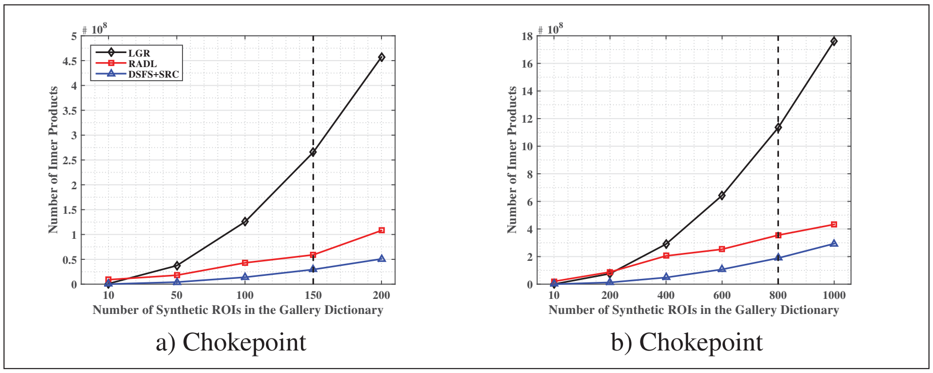
Table. 2.4 compares the complexity of the proposed DSFS-SRC algorithm with RADL (Wei & Wang (2015)), LGR (Zhu *et al.* (2014)), and flow-based frontalization (Hassner *et al.* (2015)) techniques on Chokepoint and COX-S2V datasets per each iteration. The experiments are conducted in MATLAB R2016b (64bit) Linux version on a PC workstation with an INTEL CPU (3.41-GHz) and 16GB RAM.



a) Chokepoint   b) Chokepoint

c) COX-S2V   d) COX-S2V

Figure 2.14 Average AUC and AUPR accuracy obtained by increasing the size of the generic set in the (synthetic) variant dictionary on Chokepoint (a,b) and COX-S2V (c,d) datasets.



a) Chokepoint   b) Chokepoint

Figure 2.15 Time complexity versus the number of synthetic ROIs on Chokepoint (a) and COX-S2V (b) data.

The results show our proposed method that is a joint use of generic learning and face synthesizing achieves superior recognition results compared to the other methods under the same configuration which verifies that our face synthesizing technique better preserves identity information. Although RADL, LGR, and the flow-based face frontalization techniques can achieve

Table 2.4 Average computational complexity of the DSFS and state-of-the-art methods on Chokepoint and COX datasets.

Technique	Chokepoint database		COX-S2V databasel	
	No.inner products	Run time(s)	No.inner products	Run time(s)
RADL	52,650,000	3.12	350,500,000	6.17
LGR	273,100,000	7.13	1,147,200,000	11.35
Face Frontalization+SRC	40,510,000	3.14	245,320,000	4.08
Proposed DSFS+SRC	**32,450,000**	**1.55**	**190,250,000**	**2.61**

comparable accuracy to our approach, they are computationally expensive. It can be concluded that augmenting the SRC with synthetic ROIs generated by DSFS technique has a good recognition rate with less computational cost than other state-of-the-art methods. The main reason is that the dictionary designed by DSFS technique is able to represent real-world capture conditions and does not require any traditional dictionary learning process.

2.7 Conclusions

This paper proposes a domain-specific face synthesizing (DSFS) technique to improve the performance of still-to-video FR systems when surveillance videos are captured under various uncontrolled conditions, and individuals are recognized based on a single facial image. The proposed approach takes advantage of target domain information from the generic set that can effectively represent probe ROIs. A compact set of synthetic faces is generated that resemble individuals of interest under capture conditions relevant to the target domain. For proof-of-concept validation, an augmented dictionary with a block structure is designed based on DSFS, and face classification is performed within a SRC framework. Our experiments on the Chokepoint and COX-S2V datasets show that augmenting the reference discretionary of still-to-video FR systems using the proposed DSFS approach can provide a higher level of accuracy compared to state-of-the-art approaches. The results indicated that face synthesis alone (without recovering the target domain information) cannot effectively resolve the challenges of the SSPP and visual domain shift problems. With DSFS, generic learning and face synthesis operate complementarity. The proposed DSFS technique could be improved to generate synthetic faces with expression variations for a robust FR. In addition, to improve performance, the representative synthetic ROIs generated using DSFS could be applied to generate local camera-specific ROIs. DSFS is general in that synthetic ROIs could be applied to train or fine-tune a multitude of face recognition systems like deep CNNs, with information that robust models to specific target domains.

CHAPTER 3

A PAIRED SPARSE REPRESENTATION MODEL FOR ROBUST FACE RECOGNITION FROM A SINGLE SAMPLE

Fania Mokhayeri, Eric Granger

Le Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA),
École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article published in « Pattern Recognition » 2019.

Abstract

Sparse representation-based classification (SRC) has been shown to achieve a high level of accuracy in face recognition (FR). However, matching faces captured in unconstrained video against a gallery with a single reference facial still per individual typically yields low accuracy. For improved robustness to intra-class variations, SRC techniques for FR have recently been extended to incorporate variational information from an external generic set into an auxiliary dictionary. Despite their success in handling linear variations, non-linear variations (e.g., pose and expressions) between probe and reference facial images cannot be accurately reconstructed with a linear combination of images in the gallery and auxiliary dictionaries because they do not share the same type of variations. In order to account for non-linear variations due to pose, a paired sparse representation model is introduced allowing for joint use of variational information and synthetic face images. The proposed model, called *synthetic plus variational model*, reconstructs a probe image by jointly using (1) a variational dictionary and (2) a gallery dictionary augmented with a set of synthetic images generated over a wide diversity of pose angles. The augmented gallery dictionary is then encouraged to pair the same sparsity pattern with the variational dictionary for similar pose angles by solving a newly formulated simultaneous sparsity-based optimization problem. Experimental results obtained on Chokepoint and COX-S2V datasets, using different face representations, indicate that the proposed approach

can outperform state-of-the-art SRC-based methods for still-to-video FR with a single sample per person.

3.1 Introduction

Video-based face recognition (FR) has attracted a considerable amount of interest from both academia and industry due to the wide range applications as found in surveillance and security. In contrast to FR systems based on still images, an abundance of spatio-temporal information can be extracted from target domain videos to contribute in the design of discriminant still-to-video FR systems.

Sparse Representation-based Classification (SRC) techniques can provide an accurate and cost-effective solution in many video FR applications when there are a sufficient number of reference training images per each person under controlled condition (Wright *et al.* (2009); Xu *et al.* (2017)). However, single sample per person (SSPP) problems are common in video-based security and surveillance applications, as found in, e.g., biometric authentication and watch-list screening (Nourbakhsh *et al.* (2016); Dewan *et al.* (2016)). For example, still-to-video FR systems are typically designed using only one reference still image per individual in the source domain, and then faces captured with video surveillance cameras in target domain are matched against these reference stills (Bashbaghi *et al.* (2017a,b)). Additionally, when faces are captured under challenging uncontrolled conditions, they may vary considerably according to pose, illumination, occlusion, blur, scale, resolution, expression, etc. In such cases, using SRC techniques often associated with limited robustness to intra-class variations, and a lower recognition rate.

State-of-the-art approaches designed to address SSPP problems in SRC-based FR systems can be roughly divided into three categories: (1) image patching methods, where the images are partitioned into several patches (Zhu *et al.* (2014); Gao *et al.* (2015)), (2) face synthesis technique to expand the gallery dictionary (Mokhayeri *et al.* (2019a); Hu *et al.* (2018)), and (3) generic learning methods, where a genetic training set is used to leverage variational information from an auxiliary generic set of images to represent the differences between probe and gallery images (Wei & Wang (2015); Deng *et al.* (2018)). Indeed, similar intra-class variations may be shared by different individuals in the generic set and ROIs in the gallery. Moreover, a generic set can be easily collected during operations or some camera calibration process, and encode subtle knowledge on faces appearing in the operational environment. One of the pioneering techniques in generic learning is extended SRC (ESRC) (Deng *et al.* (2012)), which manually constructs an auxiliary variational dictionary from a generic set to accurately represent a probe face with unknown variations from the target domain. ESRC was subsequently generalized to employ different sparsity for identity and variational parts in sparse coefficients (Li *et al.* (2016)), and to learn the variational dictionary that accounts for the relationship between the reference gallery and external generic set (Yang *et al.* (2013)).

Although leveraging intra-class variations from a generic set has been shown to improve robustness to some linear facial variations, it cannot accurately address non-linear facial variations (e.g., pose and expression) between reference still ROIs in the source domain and probe videos ROIs captured in real-world capture conditions in the target domain. Indeed, non-linear variations are not additive nor sharable. For instance, a probe video ROI with various lighting can be recovered with a linear combination of an image with a natural lighting and its corresponding illumination component. However, a probe ROI with a profile view cannot be accurately reconstructed with a linear combination of frontal view ROIs in gallery dictionary and profile view ROIs in the auxiliary dictionary because they do not share the same type of variations. Non-linear facial variations between still and video ROIs make it difficult to represent a probe image using a linear combination of reference and generic set images. Another concern with ESRC is the large manually designed auxiliary dictionary (obtained via random selection in the generic set) which is computationally expensive. To address these concerns, we focus on two issues: (1) how to represent a probe image under non-linear variations with a linear combination of reference set and generic set, (2) how to design a discriminative dictionary, and (3) how to yield a robust representation with a minimum number of images.

In this paper, a paired sparse representation framework referred as the *synthetic plus variational model* (S+V) is proposed to address the problem of non-linear pose variations by increasing the range of pose variations in the gallery dictionary. Since collecting a large database with a wide variety of views is extremely expensive and time-consuming, a set of synthetic face images under representative pose are generated. As illustrated in Fig. 3.1, a probe video ROI is reconstructed using an auxiliary dictionary as well as a gallery dictionary augmented with a set of synthetic face images generated under a representative diversity of azimuth angles. The proposed sparse model not only allows probe image to be represented by the atoms of both augmented and auxiliary dictionaries, but also restricts the selected atoms to be combined with the same viewpoint, thus providing an improved representation.

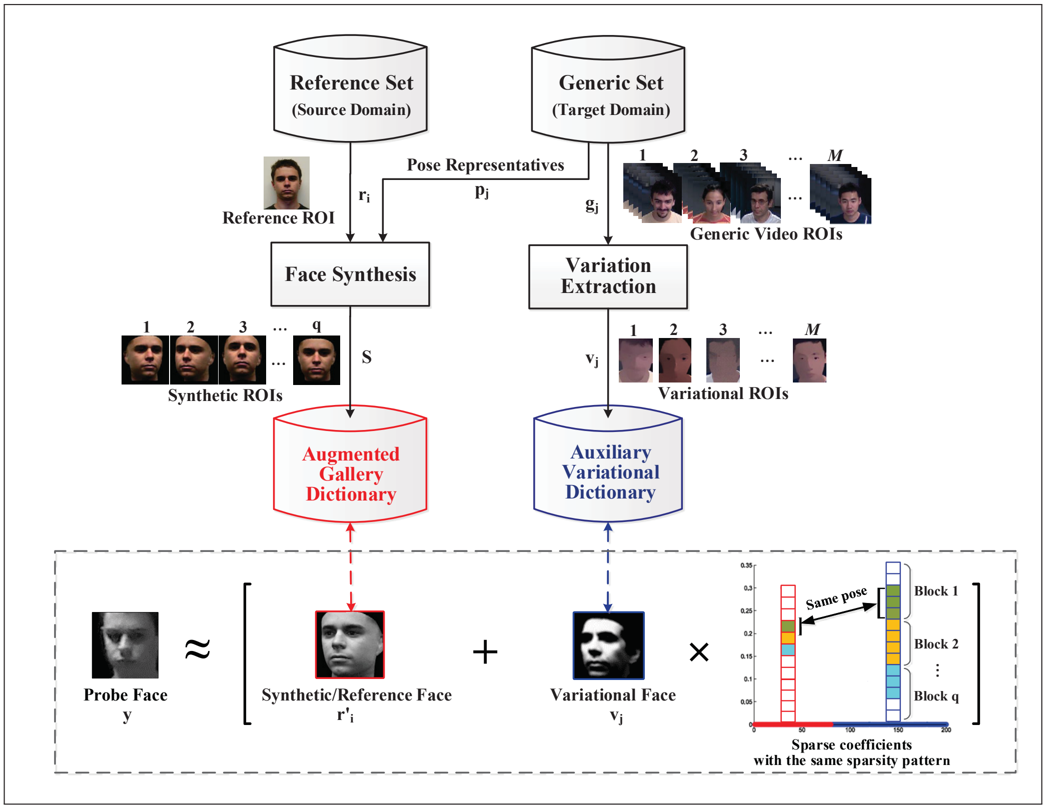


Figure 3.1 Overall architecture of the proposed approach. The gallery dictionary is augmented with a diverse set of synthetic images and the auxiliary variational dictionary co-jointly encode non-linear variations in appearance. Sparse coefficients within each dictionary share the same sparsity pattern in terms of pose angle.

Under this model, facial ROIs from trajectories in the generic set are clustered in the captured condition space (defined by pose angle) by applying row sparsity (Elhamifar *et al.* (2012)). The auxiliary variational dictionary with block structure is designed using intra-class variations as subsets of the pose clusters. Following this, the gallery dictionary is augmented with the synthetic face images generated from the original reference image in the source domain, where the rendering parameters are estimated based on the center of each cluster in the target domain. By introducing a joint sparsity structure, the pose-guided augmented gallery dictionary is encouraged to share the same sparsity pattern with the auxiliary dictionary for the same pose angles. Each synthetic facial ROI in the augmented gallery dictionary is thereby combined with approximately the same facial viewpoint in the variational dictionary in a joint manner (Rakotomamonjy (2011)). During the operation, each input probe face captured in videos is represented by a linear combination of ROIs from a same person and same pose in the augmented gallery dictionary as well as the intra-class variations from a same pose in the auxiliary variational dictionary. In this framework, the auxiliary dictionary models the linear variations (such as illumination changes, different occlusion levels) and non-linear pose variation are modeled by augmented gallery dictionary. Note that the S+V model is paired across different domains in the enrollment stage. The main contributions of this paper are:

- A generalized sparse representation model for still-to-video FR, using generic learning and data augmentation to represent both linear and non-linear variations based on only one reference still ROI;
- A simultaneous optimization technique to encourage pairing between each synthetic profile image in the augmented gallery dictionary and a similar view in the auxiliary dictionary;
- An efficient SRC method to design a compact augmented dictionary using row sparsity.

This paper extends our preliminary investigation of synthetic plus variational models (Mokhayeri & Granger (2018)) in several ways, in particular with: (1) a comprehensive analysis of dictionary design and of selection of representative face exemplars; (2) a detailed description of the proposed joint sparsity structure; and (3) more experimental results and interpretations, including results with deep facial representations, an ablation study and complexity analysis.

For proof-of-concept validation, a particular implementation of the proposed SRC technique for still-to-video FR is considered where representative pose angles are selected by applying clustering on the generic set. The original and synthetic ROIs rendered under these pose angles are employed to design an augmented gallery dictionary, while the pose clusters of video ROIs are exploited to design an auxiliary variational dictionary with block structure. The simultaneous sparsity constraint is then applied to both dictionaries to improve the discrimination power of the dictionaries. Moreover, since most state-of-the-art FR methods rely on Convolution Neural Network (CNN) architectures such as ResNet (He *et al.* (2016)) and VGGNet (Simonyan & Zisserman (2015)), the model is fed with CNN features extracted from the atoms of dictionaries (Gao *et al.* (2017); Cai *et al.* (2016)), in order to further improve still-to-video FR accuracy. Performance of the SRC implementation is evaluated on two public video FR databases – Chokepoint (Wong *et al.* (2011)) and COX-S2V (Huang *et al.* (2015)).

The rest of the paper is organized as follows. Section 3.2 provides a brief review for SRC methods that employ generic learning to address SSPP problems. Section 3.3 describes the proposed S+V model. Section 3.4 presents a particular implementation of the S+V model for still-to-video FR system. Finally, Sections 3.5 and 3.6 describe the methodology and experimental results, respectively.

3.2 Background on Sparse Coding

In the following, the set $\mathbf{D} = \{\mathbf{R}_1, \ldots, \mathbf{R}_i, \ldots, \mathbf{R}_k\} \in \mathbb{R}^{d \times N}$ denote a gallery dictionary, where $\mathbf{R}_i = \{\mathbf{r}^i_1, \ldots, \mathbf{r}^i_j, \ldots, \mathbf{r}^i_n\} \in \mathbb{R}^{d \times n}$ is composed of 1 reference still ROIs belonging to one of k different classes, d is the number of pixels or features representing a ROI and $N = kn$ is the total number of reference still ROIs. The set $\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2 \ldots, \mathbf{g}_m\} \in \mathbb{R}^{d \times m}$ denotes the auxiliary generic set composed of m external generic images of unknown persons captured in the target domain. The set $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\} \in \mathbb{R}^{d \times m}$ denotes the auxiliary variational dictionary composed of m intra-class variations extracted from $\mathbf{G} \in \mathbb{R}^{d \times m}$.

3.2.1 Sparse Representation-based Classification

Given a probe image $\mathbf{y}$, SRC represents $\mathbf{y}$ as a sparse linear combination of a reference set $\mathbf{D} \in \mathbb{R}^{d \times k}$. SRC uses the ℓ_1-minimization to regularize the representation coefficients. More precisely, SRC derives the sparse coefficient $\boldsymbol{\alpha}$ of $\mathbf{y}$ by solving the following ℓ_1-minimization problem:

$$\min_{\alpha} \|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1. \tag{3.1}$$

where λ is a regularization parameter, and $\lambda > 0$. After the sparse vector of coefficients $\boldsymbol{\alpha}$ is obtained, the probe image $\mathbf{y}$ is recognized as belonging to class k^* if it satisfies:

$$k^* = \arg\min_{k} \|\mathbf{y} - \mathbf{D}\gamma_k(\boldsymbol{\alpha})\|_2. \tag{3.2}$$

where γ_k is a vector whose only nonzero entries are the entries in $\boldsymbol{\alpha}$ that are associated with class k. SRC is based on the idea that a probe image $\mathbf{y}$ can be best linearly reconstructed by the columns of $\mathbf{D}_{k^*}$ if it belongs to class k^*. As a result, most non-zero elements of $\boldsymbol{\alpha}$ will be associated with class k^*, and $\|\mathbf{y} - \mathbf{D}\gamma_{k^*}(\boldsymbol{\alpha})\|_2$ yields the minimum reconstruction error. An important assumption of SRC is that it requires a large amount of reference training images to form an over-complete dictionary. However, in many practical applications, the number of labeled reference images are limited, and SRC accuracy declines in such cases (Wright *et al.* (2009)).

3.2.2 SRC through Generic Learning

Since the facial variations share much similarity across different individuals, an external generic set with multiple images of unknown persons as they appear in the target domain can provide discriminant information on intra-class variations. These additional variations can enrich the gallery diversity, especially in SSPP scenarios. The general model solves the following minimization problem:

$$\min_{\alpha,\beta} \left\| \mathbf{y} - [\mathbf{D}, \mathbf{V}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_a^a + \lambda \left\| \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_b^b . \tag{3.3}$$

where $\boldsymbol{\alpha}$ is a sparse vector that selects a limited number of variant bases from the gallery dictionary $\mathbf{D}$, and $\boldsymbol{\beta}$ is another sparse vector that selects a variant bases from the auxiliary variational dictionary $\mathbf{V}$, $a \in \{1,2\}$, $b \in \{1,2\}$ and $\lambda > 0$. The variant bases can be estimated by subtracting the natural (original) image of a class from other images of the same class, the difference from the class centroid, and pairwise difference. The probe image $\mathbf{y}$ is recognized as belonging to class k^* if it satisfies:

$$k^* = \arg\min_k \left\| \mathbf{y} - [\mathbf{D}, \mathbf{V}] \begin{bmatrix} \gamma_k(\boldsymbol{\alpha}) \\ \boldsymbol{\beta} \end{bmatrix} \right\|_a^a . \tag{3.4}$$

where γ_k is reused as a matrix operator.

Deng *et al.* (2012) introduced extended SRC (ESRC), which manually designs an auxiliary dictionary (through random selection from a generic set) to accurately represent a probe face with unknown variations from the target domain. The model of Eq. 3.4 degenerates to the ESRC model when $a = 2$ and $b = 1$. Motivated by ESRC, Yang *et al.* (2013) proposed the sparse variation dictionary learning (SVDL) model to learn the variational dictionary by accounting for the relationship between the reference gallery and external generic set. A robust auxiliary dictionary learning (RADL) technique was proposed by Wei & Wang (2015) that extracts representative information from external data via dictionary learning without assuming the prior knowledge of occlusion in probe images. Nourbakhsh *et al.* (2016) integrated variational information from the target domain with the reference gallery set through domain adaptation to enhance the facial models for still-to-video FR. Fan *et al.* (2018) proposed a new approach to learn a kernel SRC model based on a virtual dictionary and the original training set. Deng *et al.* (2018) developed a superposed linear representation classifier to cast the recognition problem by representing the test image in terms of a superposition of the class centroids and the shared intra-class differences. A sparse illumination and transfer learning technique proposed by Zhuang *et al.* (2015) allows fitting illumination examples of auxiliary face images from one or more additional subjects with a sparsely-used illumination dictionary. A local generic representation-based (LGR) framework for FR with SSPP was proposed by Zhu *et al.* (2014). It builds a gallery dictionary by extracting the patches from the gallery database, while an intra-class variation dictionary is formed by using an external generic set to predict the possible facial variations (*e.g.*, illuminations, pose, and expressions). In order to address non-linearity, Fan *et al.* (2018) used a nonlinear mapping to transform the original reference data into a high dimensional feature space, which is achieved using a kernel-based method. A customized SRC (CSR) had been proposed to leverage the different sparsity of identity and variational parts in sparse coefficients, and to assign different parameters to their regularization terms (Li *et al.* (2016)). Yang *et al.* (2017) presented a joint and collaborative sparse representation framework that exploits the distinctiveness and commonality of different local regions. A novel discriminative approach is proposed by Lin *et al.* (2018a), in which a robust dictionary is learned from diversities in training samples, generated by extracting and generating facial variations. Xie *et al.* (2019) proposed feature sparseness-based regularization to learns deep features with better generalization capabilities. In this paper, the regularization is integrated into the original loss function, and optimized with a deep metric learning framework. Luo *et al.* (2019) proposed a novel multi-resolution dictionary learning method for FR that provides multiple dictionaries – each one associated with a resolution – while encoding the similarity of representations obtained using different dictionaries in the training phase. 3D Morphable Model (3DMM), proposed by Blanz & Vetter (2003), has been widely used to synthesize new face images from a single 2D face image. The 3DMM is expanded by adopting a shared covariance structure to mitigate small sample estimation problems associated with data in high dimensional spaces (Koppen *et al.* (2018)). It models the global population as a mixture of Gaussian sub-populations, each with its own mean value. Finally, an efficient deep learning model for face synthesis is proposed by Jiao *et al.* (2018) which dose no rely on complex optimization.

Zhang *et al.* (2018a) proposed a deep learning model that automatically generates synthetic face images with different expressions using GAN under arbitrary poses to enlarge the training set. Unlike our representative selection technique, they embedded a classifier into the network to generate representative face images.

The aforementioned techniques work well in video-based FR. However, they neglect the impact of non-linear variations between probe images and facial images in the gallery and auxiliary dictionaries. To account for the non-linearities, particularly pose variations, the range of view-points represented in the gallery dictionary should be increased to represent the probe image with the same view gallery and variations, and thereby compensate the non-linear pose variations. Additionally, the sparsity pattern should ensure the correlation between the gallery and variational dictionaries in terms of pose angles.

3.3 The Proposed Approach - A Synthetic plus Variational Model

In this section, a new sparse representation model – called the *Synthetic plus Variational* (S+V) model – is proposed to overcome issues related to the non-linear pose variations with conventional and ESRC model. SRC techniques commonly assumed that frontal and profile views share the same type of variations. To address this limitation, we increase the range of pose variations of gallery dictionary to represent the probe with the same view gallery and variations, and accordingly compensate the non-linear pose variations.

The proposed S+V model exploits two dictionaries including (1) an augmented gallery dictionary containing the original reference still ROI of each individual as well as their synthetic profile ROIs (with diverse poses) enrolled to the still-to-video FR system, and (2) an auxiliary variational dictionary which contains variations from the target domain that can be shared by different persons. Two dictionaries are correlated by imposing the simultaneous sparsity prior that force the augmented gallery dictionary to pair the same sparsity pattern with the auxiliary dictionary for the same pose angles. In this manner, each synthetic profile image in the augmented gallery dictionary is combined with approximately the similar view in the auxiliary dictionary. Fig. 3.2 gives an illustrative example that compares the sparsity structure of SRC, ESRC and S+V model. The rest of this section presents more details on the dictionary design and encoding process with the S+V model.

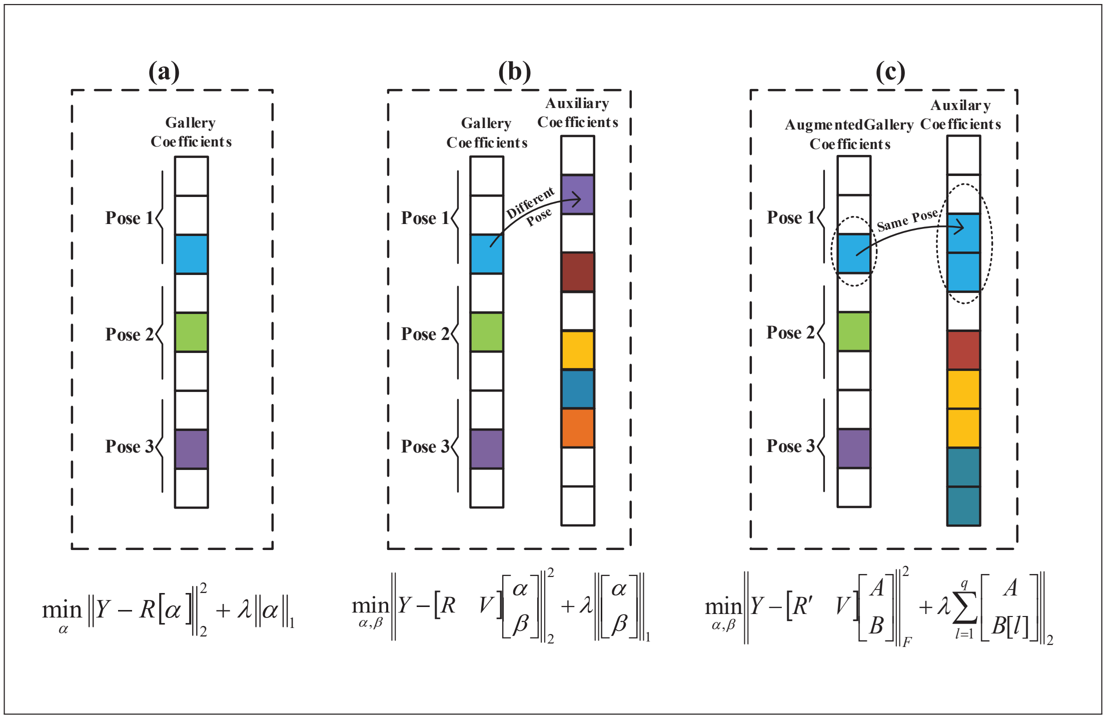

(a)
Gallery Coefficients
Pose 1
Pose 2
Pose 3
$\min_{\alpha} \|Y - R[\alpha]\|_2^2 + \lambda\|\alpha\|_1$
(b)
Gallery Coefficients
Auxiliary Coefficients
Different Pose
Pose 1
Pose 2
Pose 3
$\min_{\alpha,\beta} \left\| Y - [R \quad V] \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_2^2 + \lambda \left\| \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \right\|_1$
(c)
AugmentedGallery Coefficients
Auxilary Coefficients
Same Pose
Pose 1
Pose 2
Pose 3
$\min_{\alpha,\beta} \left\| Y - [R' \quad V] \begin{bmatrix} A \\ B \end{bmatrix} \right\|_F^2 + \lambda \sum_{l=1}^{q} \left\| \begin{bmatrix} A \\ B[l] \end{bmatrix} \right\|_2$

Figure 3.2 A comparison of the coefficient matrices for three sparsity models: (a) Independent sparsity (SRC) with a single dictionary, (b) Extended sparsity (ESRC) with two dictionaries, and (c) Paired extended sparsity (S+V model) with pair-wise correlation between two dictionaries where the sparse coefficients of same poses share the same sparsity pattern. Each column represents a sparse coefficient vector and each square block denotes a coefficient value. White blocks denote zero values, whereas color blocks stand for nonzero values.

3.3.1 Dictionary Design

In order to design the gallery and auxiliary dictionaries, the representative pose angles are determined by characterizing the capture conditions from a large generic set of video ROIs in the pose space (estimations of pitch, roll, and yaw). Prior to operation, e.g., during a camera calibration process, facial ROIs are isolated in facial trajectories from the videos of unknown persons captured in the target domain. A representative set of video ROIs are selected by applying row sparsity regularized optimization program on facial trajectories in the captured condition space defined by pose angles. Next, the variational information of the generic set with multi-samples per person are extracted to form an auxiliary dictionary based on the subsets of the pose clusters. A compact set of synthetic images is then generated from the reference set in the source domain based on the information obtained from the center of each cluster in the target domain, called pose representatives, and integrated into the gallery dictionary to enrich the diversity of the gallery set. Two dictionaries are correlated by imposing the simultaneous sparsity prior that force the same sparsity patterns among the multiple sparse representation vectors in the augmented and auxiliary dictionaries in terms of pose angles. Finding representative poses not only are employed to make a pair-wise correlation between the dictionaries but also can save time and memory and improve the recognition performance due to preventing over-fitting. Inspired by Elhamifar *et al.* (2012); Elhamifar & Kaluza (2017), we formulated the representative selection problem as a row sparsity regularized trace minimization problem where the objective is to find a few representatives (exemplars) that efficiently represent the collection of data points according to their dissimilarities.

The proposed model allows to select pose representatives from a collection of N pose samples. The pose angles are estimated using the discriminative response map fitting method (Asthana *et al.* (2013)) which is a state-of-the-art method for accurate fitting, suitable for handling occlusions and changing illumination conditions. The estimated head pose for the j^{th} video ROI ($\mathbf{g}_j$) in the generic set is defined as $\boldsymbol{\theta}_j = (\boldsymbol{\theta}_j^{pitch}, \boldsymbol{\theta}_j^{yaw}, \boldsymbol{\theta}_j^{roll})$. Euler angles $\boldsymbol{\theta}^{pitch}$, $\boldsymbol{\theta}^{yaw}$, and $\boldsymbol{\theta}^{roll}$ are used to represent roll, yaw and pitch rotation around X axis, Y axis, and Z axis of the global coordinate system, respectively. The set of dissimilarities $\{d_{ij} : i, j = 1, ..., k\}$ between every pair of pose data points are then calculated by using the Euclidean distance, which indicates how well the data point i is suited to be an exemplar of data point j. The dissimilarities are arranged into matrix:

$$\mathbf{D} \triangleq \begin{bmatrix} \mathbf{d}_1^T \\ \vdots \\ \mathbf{d}_N^T \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ d_{k1} & d_{k2} & \cdots & d_{kk} \end{bmatrix} \in \mathbb{R}^{k \times k}, \tag{3.5}$$

where $\mathbf{d}_i$ denotes the i^{th} row of $\mathbf{D}$. Variables z_{ij} are associated with dissimilarities d_{ij}, and organized into matrix of the same size as:

$$\mathbf{Z} \triangleq \begin{bmatrix} \mathbf{z}_1^T \\ \vdots \\ \mathbf{z}_N^T \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{kk} \end{bmatrix} \in \mathbb{R}^{k \times k}, \tag{3.6}$$

where $z_i \in \mathbb{R}^k$ denotes the i^{th} row of $\mathbf{z}$. z_{ij} is the probability that data point i is representative for data point j, and $z_{ij} \in [0,1]$. The row sparsity regularized trace minimization algorithm is applied on matrix $\mathbf{Z}$ to select some representative exemplars that can suitably encode pose data according to dissimilarities as follows:

$$\min \sum_{j=1}^{k} \sum_{i=1}^{k} d_{ij} z_{ij} + \eta \sum_{i=1}^{k} \left\| z_i \right\|_q, \tag{3.7}$$

subject to:

$$z_{ij} \geq 0, \quad \forall i,j; \quad \sum_{i=1}^{k} z_{ij} = 1, \quad \forall j, \tag{3.8}$$

where the parameter $\eta > 0$, rank regularization parameter, sets the trade-off between these two terms. As we change η in 3.7, the number of representatives found by the algorithm changes. For small values of η, where we put more emphasis on better encoding data points via representatives, we obtain more representatives. On the other hand, for large values of η, where we put more emphasis on the row sparsity of $\mathbf{Z}$, we select a small number of representatives.

Once this optimization problem (Eq. 3.7) has been solved, one can find the representative indices from the nonzero rows of $\mathbf{Z}$. The clustering of data points into K clusters, associated with K representatives, is obtained by assigning each data point to its closest representative. In particular, if { $i_1; \ldots ; i_q$ } denote the indices of the representatives, data point j is assigned to the pose representative $\theta(j)$ such that $\theta(j) = \arg\min_{\ell \in \{i_1;\ldots;i_q\}} d_{\ell j}$.

The auxiliary dictionary is designed based on these pose clusters, where each cluster forms a block in the dictionary. The pose angle of representative video ROI of each pose cluster, referred as pose exemplar, is used as rendering parameter to generate synthetic face images with varying poses using off-the-shelf 3D face models (Blanz & Vetter (2003); Tran *et al.* (2017a,b)). In this way, q synthetic profile faces, $\mathbf{S} = \{\mathbf{S}_i : i = 1, \dots, k\}$, are generated under the representative pose angles from a given single still face image where $\mathbf{S}_i = \{\mathbf{s}^i_1, \mathbf{s}^i_2, \dots, \mathbf{s}^i_q\} \in \mathbb{R}^{d \times q}$.

The augmented gallery dictionary $\mathbf{D}' = \{\mathbf{D}'_i : i = 1, \dots, k\}$, is formed by merging each still ROI of reference set with q synthetic images rendered w.r.t. representative pose exemplars, where here $\mathbf{D}'_i = \{\mathbf{r}_1, \mathbf{s}^i_1, \mathbf{s}^i_2, \dots, \mathbf{s}^i_q\} \in \mathbb{R}^{d \times (1+q)}$.

3.3.2 Synthetic Plus Variational Encoding

With the S+V model (see Fig. 3.3), each probe video ROI is seen as a combination of two different sub-signals in the augmented gallery dictionary and auxiliary variation dictionary in the linear additive model:

$$\mathbf{y} = \mathbf{D}'\boldsymbol{\alpha} + \mathbf{V}\boldsymbol{\beta} + e, \tag{3.9}$$

where $\mathbf{D}' \in \mathbb{R}^{d \times k(q+1)}$ denote the augmented gallery dictionary, $\mathbf{V} \in \mathbb{R}^{d \times m}$ denote the variational dictionary, and e is a noise term. This model searches for the sparsest representation of the probe sample in both $\mathbf{D}'$ and $\mathbf{V}$ dictionaries. We first extend the original ESRC to the following robust formulation (Eq. 3.10).

$$\min_{\alpha,\beta} \left\| \mathbf{y} - [\mathbf{D}', \mathbf{V}] \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 + \mu \|\boldsymbol{\beta}\|_\tau, \tag{3.10}$$

where $\|\cdot\|_\tau$ corresponds with combination of Gaussian and Laplacian priors, defined as Eq. 3.11. This model assigns different regularization parameters to the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ coefficients to guaranty the robustness of the variational information from generic set (Li *et al.* (2016)).

$$\|x\|_\tau = \tau\|x\|_1 + (1-\tau)\|x\|_2. \tag{3.11}$$

The simultaneous sparsity constraint is then imposed to fully benefit from the variational information as well as synthetic still ROIs. Each generic set cluster found during the representative selection forms a block in the auxiliary dictionary, and exemplar of each cluster is considered as rendering parameter in face synthesizing for augmenting the gallery dictionary. The same sparsity pattern constraint in terms of the pose angle is imposed on the dictionaries which encourages similar pose angles to select the same set of atoms for representing each view. In this way, the coefficient vectors for the still ROIs in the augmented gallery dictionary are forced to share the same sparsity pattern with non-zero coefficients associated with the video ROI belonging to the corresponding block (cluster) of the same view in the auxiliary dictionary. This improves the discrimination power of the dictionaries accordingly. The new sparse

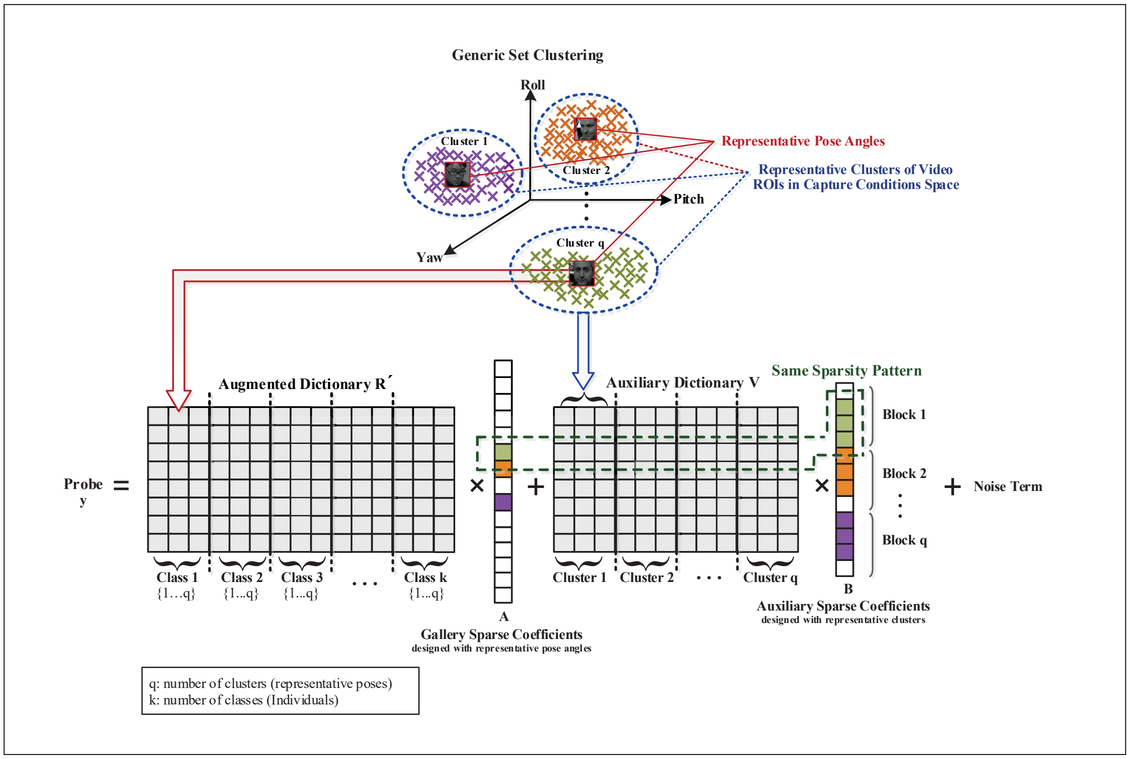

Generic Set Clustering
Roll
Pitch
Yaw
Cluster 1
Cluster 2
Cluster q
Representative Pose Angles
Representative Clusters of Video ROIs in Capture Conditions Space
Augmented Dictionary R´
Auxiliary Dictionary V
Same Sparsity Pattern
Probe y
=
×
+
×
+
Noise Term
Class 1 {1...q}
Class 2 {1...q}
Class 3 {1...q}
Class k {1...q}
Cluster 1
Cluster 2
Cluster q
Block 1
Block 2
Block q
A
Gallery Sparse Coefficients
designed with representative pose angles
B
Auxiliary Sparse Coefficients
designed with representative clusters
q: number of clusters (representative poses)
k: number of classes (Individuals)

Figure 3.3 An illustration of sparsity pattern with the S+V model based on clustering results in the pose space. Each column represents a sparse representation vector, each square denotes a coefficient and each matrix is a dictionary.

coefficients can be obtained by solving the following optimization problem:

$$\min_{A,B} \left\| \mathbf{y} - [\mathbf{D}', \mathbf{V}] \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \right\|_F^2 + \lambda \left\| \mathbf{A} \right\|_1 + \mu \sum_{l=1}^{q} \left\| \mathbf{B}[l] \right\|_\tau, \tag{3.12}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\mathbf{A} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{k(q+1)}]$ and $\mathbf{B} = [\boldsymbol{\beta}[1], \boldsymbol{\beta}[2], \ldots, \boldsymbol{\beta}[q]]$ are coefficients matrix consists of q blocks which q is number of clusters/representatives.

$$\begin{bmatrix} \widehat{\mathbf{A}} \\ \widehat{\mathbf{B}} \end{bmatrix} = \arg\min \left\| \mathbf{y} - [\mathbf{D}', \mathbf{V}] \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \right\|_F^2 + \lambda \left\| \mathbf{A} \right\|_1 + \mu \sum_{l=1}^{q} \left\| \mathbf{B}[l] \right\|_\tau, \tag{3.13}$$

subject to:

$$\left\| \mathbf{A}, \mathbf{B} \right\|_{2,1} \leq \xi, \tag{3.14}$$

where ξ is the sparsity level and $\|\cdot\|_{2,1}$ is the mixed norm defined as the sum of $\ell_2 - norm$ of all rows of matrix $\mathbf{A}$ and $\mathbf{B}$ and then applying $\ell_1 - norm$ on the obtained vector. Note that each view in formulation of Eq. 3.13 shares the same sparsity pattern at class-level, but not necessarily at atom-level in real world scenarios. This problem, called joint dynamic sparse representation, can be solved by applying $\ell_0 - norm$ across the $\ell_2 - norm$ of the sparse coefficients as follows:

$$\begin{bmatrix} \widehat{\mathbf{A}} \\ \widehat{\mathbf{B}} \end{bmatrix} = \arg\min \left\| \mathbf{y} - [\mathbf{D}', \mathbf{V}] \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \right\|_F^2 + \lambda \left\| \mathbf{A} \right\|_1 + \mu \sum_{l=1}^{q} \left\| \mathbf{B}[l] \right\|_\tau, \tag{3.15}$$

subject to:

$$\left\| \mathbf{A}, \mathbf{B} \right\|_G \leq \xi, \tag{3.16}$$

where $\left\| \cdot \right\|_G$ is defined as follows:

$$\left\| \mathbf{A}, \mathbf{B} \right\|_G = \left\| \left[\left\| \mathbf{A}_{g_1}, \mathbf{B}_{g_1} \right\|_2, \left\| \mathbf{A}_{g2}, \mathbf{B}_{g2} \right\|_2, \ldots \right] \right\|_0. \tag{3.17}$$

The use of joint dynamic sparsity regularization term allows combining the cues from all the views during joint sparse representation. Moreover, it provides a better representation of the multiple view images, which represent different measurements of the same individual from different viewpoints. Finally, the residuals for each class k are calculated for the final classification as follows:

$$r_k(y) = \left\| \mathbf{y} - [\mathbf{D}', \mathbf{V}] \begin{bmatrix} \gamma_k(\widehat{\mathbf{A}}_k) \\ \widehat{\mathbf{B}}_k \end{bmatrix} \right\|_F^2, \tag{3.18}$$

where γ_k is a vector whose nonzero entries are the entries in $\widehat{\mathbf{A}}_k$ that are associated with class k. Then the class with the minimum reconstruction error is regarded as the label for the probe subject y. Algorithm 3.1 summarizes the S+V model for still-to-video FR from a SSPP.

Algorithm 3.1 Synthetic Plus Variational Model.

```
Input: Reference still ROIs $\mathbf{D} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_k\} \in \mathbb{R}^{d \times k}$, Generic set
    $\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_m\} \in \mathbb{R}^{d \times m}$, probe sample $\mathbf{y}$, and parameters $\lambda$, $\mu$, and $\xi$.
Estimate pose angles of $\mathbf{G}$.
Apply row sparsity clustering in the pose space of $\mathbf{G}$, and produce $q$ clusters (representative
  exemplars).
Find center of each cluster as $q$ representative pose angles.
Construct the variation dictionary, $\mathbf{V} \in \mathbb{R}^{d \times m}$, with $q$ blocks.
for each $\mathbf{r}_i$ do
    Generate $q$ synthetic images $\mathbf{S}_i \in \mathbb{R}^{d \times q}$ per each individual based on $q$ representative pose
      angle.
    Merge $\mathbf{S}_i$ with $\mathbf{r}_i$ to form $\mathbf{D}'_i \in \mathbb{R}^{d \times (1+q)}$.
end
Solve the sparse representation problem to estimate coefficient matrix, $\mathbf{A}$ and $\mathbf{B}$, for $y$ by
  Eq. 3.12.
Compute the residual, $r_k(y)$ by Eq. 3.18.
Output: $label(y) = \arg\min_k (r_k(y))$.
```

3.4 Face Recognition with the S+V Model

In this section, a particular implementation is considered (see Fig. 3.4) to assess the impact of using the S+V model for still-to-video FR. The augmented and auxiliary dictionaries are constructed by employing the representative synthetic ROIs and generic variations, respectively, and classification is performed by SRC while the generic set in the auxiliary dictionary is forced to combine with approximately the same facial viewpoint in the augmented gallery dictionary. The main steps of the proposed domain-invariant FR with the S+V model are summarized as follows.

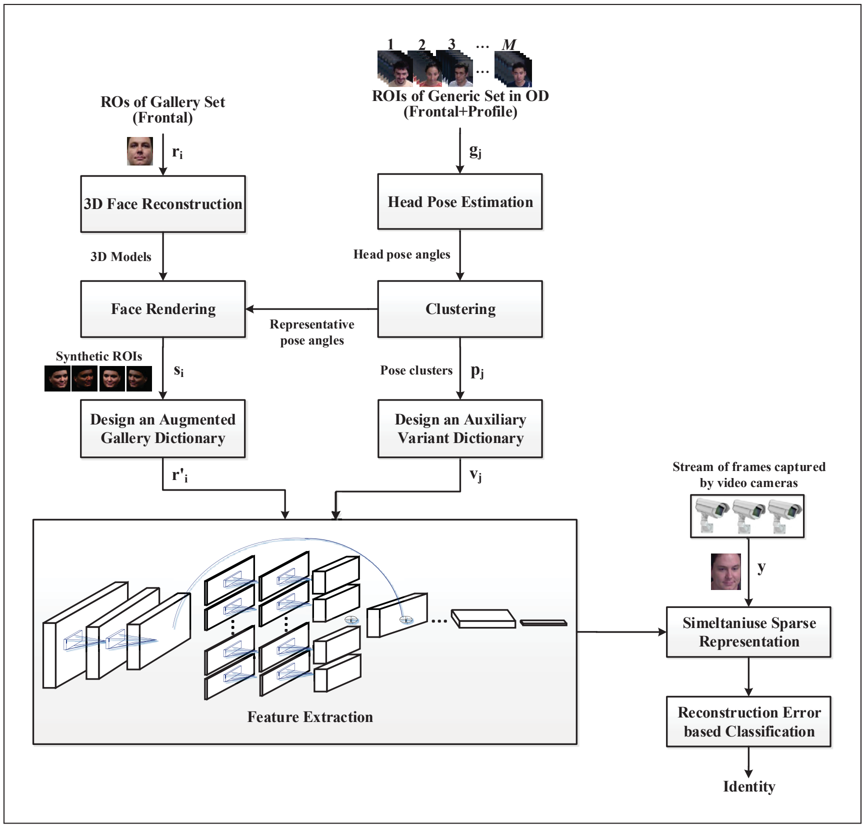

1 2 3 ... M
ROIs of Generic Set in OD (Frontal+Profile)
ROs of Gallery Set (Frontal)
r_i
g_j
3D Face Reconstruction
Head Pose Estimation
3D Models
Head pose angles
Face Rendering
Clustering
Representative pose angles
Synthetic ROIs
s_i
Pose clusters
p_j
Design an Augmented Gallery Dictionary
Design an Auxiliary Variant Dictionary
r'_i
v_j
Stream of frames captured by video cameras
y
Simeltaniuse Sparse Representation
Feature Extraction
Reconstruction Error based Classification
Identity

Figure 3.4 Block diagram of the proposed still-to-video FR system with the S+V modeling.

- **Step 1.** Select Representatives: The generic set $\mathbf{G}_i \in R^{d \times m}$ in the target domain is clustered based on their pose angles based on row sparsity.
- **Step 2.** Design an Augmented Gallery Dictionary: The q synthetic ROIs $\mathbf{S}_i \in R^{d \times q}$ are generated for each $\mathbf{r}_i$ of the reference gallery set in the source domain to form an augmented gallery dictionary $\mathbf{D}'_i \in \mathbb{R}^{d \times k(q+1)}$, where q is the number of clusters/representatives.
- **Step 3.** Form an Auxiliary Dictionary: The variations of the natural albedo of the generic set $\mathbf{G}_i \in R^{d \times m}$ in the target domain are extracted by subtracting the natural image from

other images of the same class to form a generic auxiliary dictionary $\mathbf{V}_i \in R^{d\times m}$ with block structure.

- **Step 4.** Extract Features: The deep CNN features of $\mathbf{D}'_i \in \mathbb{R}^{d\times k(q+1)}$ and $\mathbf{V}_i \in R^{d\times m}$ are extracted.
- **Step 5.** Apply Simultaneous Sparsity: The augmented gallery dictionary is encouraged to pair the sparsity pattern with the auxiliary dictionary for the same pose angles by applying the simultaneous sparsity.
- **Step 6.** Validation: The proposed system assess if given probe ROIs belong to one of the enrolled persons and rejects invalid probe ROIs using *sparsity concentration index (SCI)* criteria defined by Wright *et al.* (2009):

$$\mathrm{SCI}(\hat{\mathbf{x}}) \doteq \frac{k.\max_i \parallel \delta_i(\hat{\mathbf{x}}) \parallel_1 / \parallel \hat{\mathbf{x}} \parallel_1 -1}{k-1} \quad \in [0,1]\ . \tag{3.19}$$

A probe ROI is accepted as valid if $SCI \geq \tau$ and otherwise rejected as invalid, where $\tau \in (0,1)$ is an outlier rejection threshold.

3.5 Experimental Methodology

3.5.1 Datasets

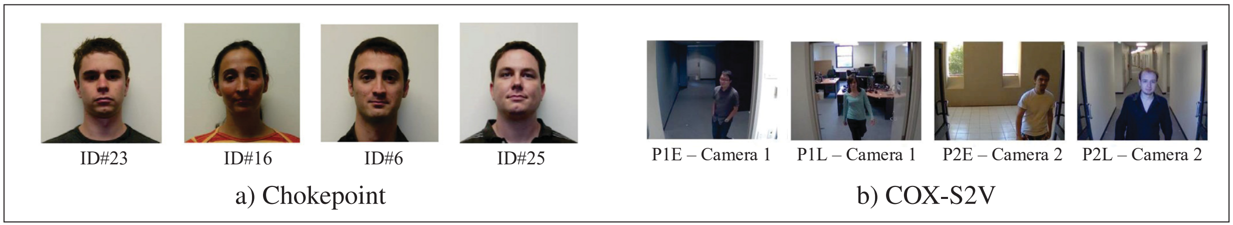
In order to evaluate the performance of the proposed S+V model for still-to-video FR, an extensive series of experiments are conducted on Chokepoint[1] (Wong *et al.* (2011)) and COX-S2V[2] (Huang *et al.* (2015)) datasets. Chokepoint and COX-S2V datasets are suitable for experiments in still-to-video FR in video surveillance because they are composed of a high-quality still image and lower-resolution video sequences, with variations of illumination conditions, pose, expression, blur and scale.

[1] http://arma.sourceforge.net/chokepoint.

[2] http://vipl.ict.ac.cn.

Chokepoint (Wong *et al.* (2011)) (see Fig. 3.5) consists of 25 subjects walking through portal 1 (P1) and 29 subjects in portal 2 (P2). Videos are recorded over 4 sessions (S1,S2,S3,S4) one month apart. An array of 3 cameras (Cam1,Cam2,Cam3) are mounted above P1 and P2 that capture the subjects during 4 sessions while they are either entering (E) or leaving (L) the portals in a natural manner. In total, 4 data subsets are available (P1E, P1L, P2E, and P2L), and the dataset consists of 54 video sequences.

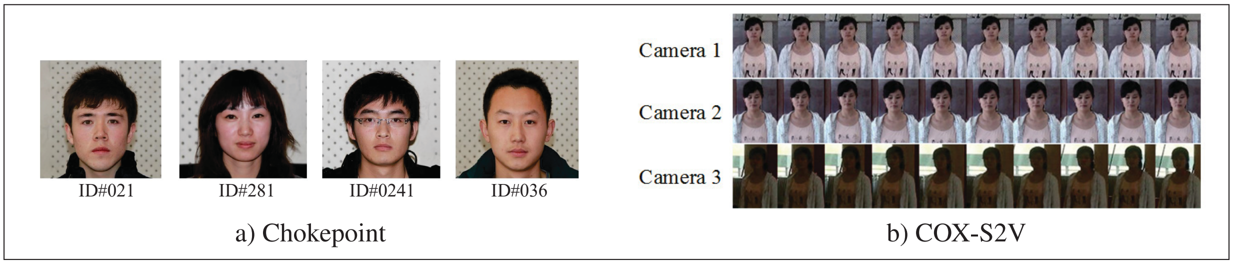
COX-S2V dataset (Huang *et al.* (2015)) (see Fig. 3.6) contains $1,000$ individuals, with 1 high-quality still image and $3,000$ low-resolution video sequences per each individual simulating video surveillance scenario. The video frames are captured by 4 cameras (Cam1, Cam2, Cam3, Cam4) mounted at fixed locations of about 2 meters high. In each video, an individual walk through an S-shape route with changes in pose, illumination, and scale.



a) Chokepoint

b) COX-S2V

Figure 3.5 Examples of still images and video frames from portals and cameras of Chokepoint dataset.



a) Chokepoint

b) COX-S2V

Figure 3.6 Examples of still images and video frames from 3 cameras of COX-S2V dataset.

3.5.2 Protocol and Performance Measures

A particular implementation of the S+V model for still-to-video FR has been considered to validate the proposed approach. We hypothesize that accuracy can be improved by adding synthetic reference faces to the gallery dictionary and encouraging the dictionaries to share the same sparsity pattern for the same pose angles can address non-linear pose variations.

First, it is assumed that during the calibration process, q representative pose angles are selected based on the q pose clusters obtained from facial ROI trajectories of unknown persons captured in the target domain using the row sparsity clustering. During the enrollment of an individual to the system, q synthetic ROIs for each reference still ROI are generated under typical pose variations from different camera viewpoints. For face synthesis, we employ the conventional 3D Morphable Model (3DMM) (Blanz & Vetter (2003)) and the CNN-regressed 3DMM (Tran *et al.* (2017a)), that relies on a CNN for regressing 3DMM parameters. The gallery dictionary is constructed using the reference still ROIs of the individuals along with their synthetic ROIs. Next, the auxiliary variational dictionary is designed using the intra-class variations of the generic set with block structure (q blocks). Additionally, we consider extracting deep features using CNN models to further improve the FR recognition rate. The networks are pre-trained using the VGGFace2 dataset with AlexNet (Krizhevsky *et al.* (2012)), ResNet (He *et al.* (2016)) and VGGNet (Simonyan & Zisserman (2015)) architectures using Triplet Loss (Schroff *et al.* (2015)). The extracted features are concatenated as a row feature vector of this dictionary. The sparse model is fed with the extracted features. In all experiments with Chokepoint dataset, 5 target individuals are selected randomly to design a watch-list that includes a high-quality frontal captured images, and for the experiment with COX-S2V, 20 individuals are randomly selected to build a watch-list from high-quality faces. Videos of 10 individuals that are assumed to come from non-target persons are used as generic set. The rest of the videos including 10 other non-target individuals and the videos of individuals who are already enrolled in the watch-list are used for testing. In order to obtain representative results, this process is repeated 5 times with a different random selection of watch-lists and the average performance is reported with standard deviation over all the runs.

During the operational phase, FR is performed by sparse coding the features of probe ROI over the features of augmented and auxiliary (variational) dictionaries ROIs. The sparsity parameter λ is fixed to 0.005 during the experiments. We also compared the S+V method to several baseline state-of-the-art methods: ESRC (Deng *et al.* (2012)), SVDL (Yang *et al.* (2013)), RADL (Wei & Wang (2015)), LGR (Zhu *et al.* (2014)), CSR (Li *et al.* (2016)), face frontalization (Hassner *et al.* (2015)), and recognition via generation (Masi *et al.* (2016)).

The average performance of the proposed and baseline FR systems is measured in terms of accuracy and complexity. For accuracy, we measure the partial area under ROC curve pAUC(20%) (using the AUC at $0 < FPR \leq 20\%$) and area under precision-recall space (AUPR). An estimation of time complexity is provided analytically based on the worst-case number of operations performed per iteration. Then, the average running time of our algorithm is measured with a randomly selected probe ROIs using a PC workstation with an Intel Core i7 CPU (3.41GHz) processor and 16GB RAM.

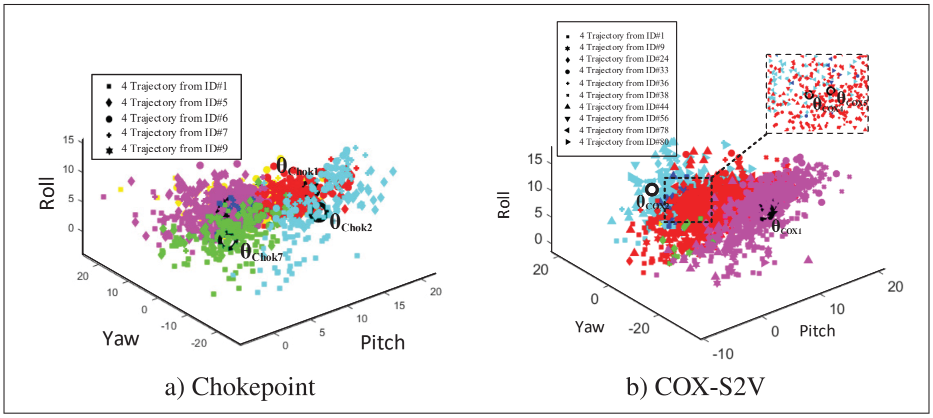
3.6 Results and Discussion

This section first shows some examples of synthetic face images produced under representative pose variations, and then presents still-to-video FR performance achieved with augmenting SRC dictionaries with such images to address non-linear variations caused by pose changes. In order to investigate the impact of the proposed S+V model on performance, we considered the still-to-video FR system described in Section 3.4 with a growing number of synthetic faces, along with a generic training set. Finally, this section presents an ablation study (showing the effect of each module on the performance) and a complexity analysis for our proposed approach.

3.6.1 Synthetic Face Generation

Fig. 3.7 shows an example of the clustering (based on row sparsity) obtained with facial ROIs of 20 trajectories extracted from Chokepoint videos of 5 individuals and 40 trajectories extracted from COX-S2V videos of 10 individuals in the 3-dimensional pose (roll-pitch-yaw) space. In this experiment, $q_{Chok} = 7$ and $q_{COX} = 6$ representative pose condition clusters are typically determined using row sparsity with Chokepoint and COX-S2V data, respectively. The exemplars selected from these clusters (black circles) are used to define representative pose angles for synthetic face generation with 3DMM and 3DMM-CNN techniques. For instance, the representative pose angles with the Chokepoint database, are listed as follows: $\theta_{Chok1} =$ (pitch, yaw, roll)= (15.65, 14.77, -0.62), $\theta_{Chok2} =$ (12.44, 2.76, 3.64), $\theta_{Chok3} =$ (9.06, -5.46, 4.73), $\theta_{Chok4} =$ (1.98, 6.09, 2.79), $\theta_{Chok5} =$ (13.21, 15.32, 6.14), $\theta_{Chok6} =$ (0.64, -18.93, 0.86), $\theta_{Chok7} =$ (5.23, 2.92, 2.03) degrees.

Fig. 3.8 and 3.9 show the synthetic face images generated based on 3DMM and 3DMM-CNN under representative exemplars using reference still ROIs of the Chokepoint and COX-S2V datasets, respectively.



a) Chokepoint

b) COX-S2V

Figure 3.7 Example of clusters obtained with 20 and 40 facial trajectories represented in the pose space with Chokepoint (ID#1, #5, #6, #7, #9) and COX-S2V (ID#1, #9, #24, #33, #36, #38, #44, #56, #78, #80) datasets, respectively. Clusters are shown with different colors, and their representative pose exemplars are indicated with a black circle.

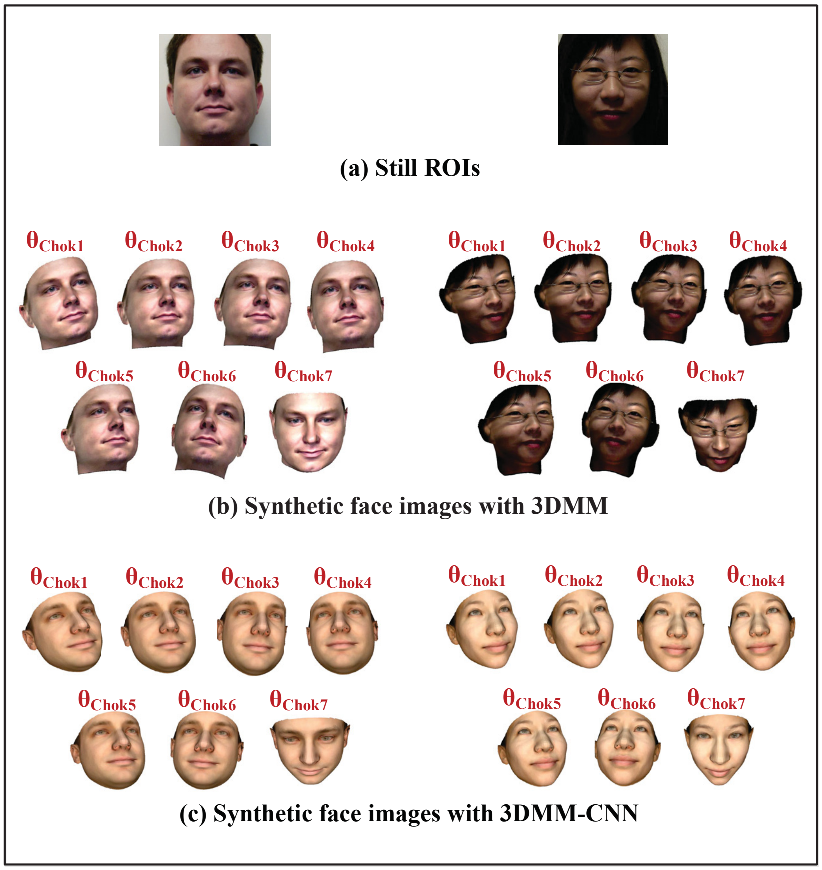


Figure 3.8 Examples of synthetic face images generated from the reference still ROI of individuals ID#25 and ID#26 (a) of Chokepoint dataset. They are produced based on representative exemplars (poses) and using 3DMM (b) and 3DMM-CNN (c).

3.6.2 Impact of Number of Synthetic Images

In this subsection, the proposed S+V model is evaluated for a growing set of synthetic facial images in the augmented gallery dictionary. Fig. 3.10 shows the average pAUC(20%) and AUPR accuracy obtained for the implementation in Section 3.4 when increasing the number of synthetic ROIs per each individual. These ROIs were sampled from the q representative pose exemplars from the Chokepoint and COX-S2V datasets. Results indicate that adding representative synthetic ROIs to the gallery dictionary allows to outperform the baseline system designed with an original reference still ROI alone. AUC and AUPR accuracy increase con-

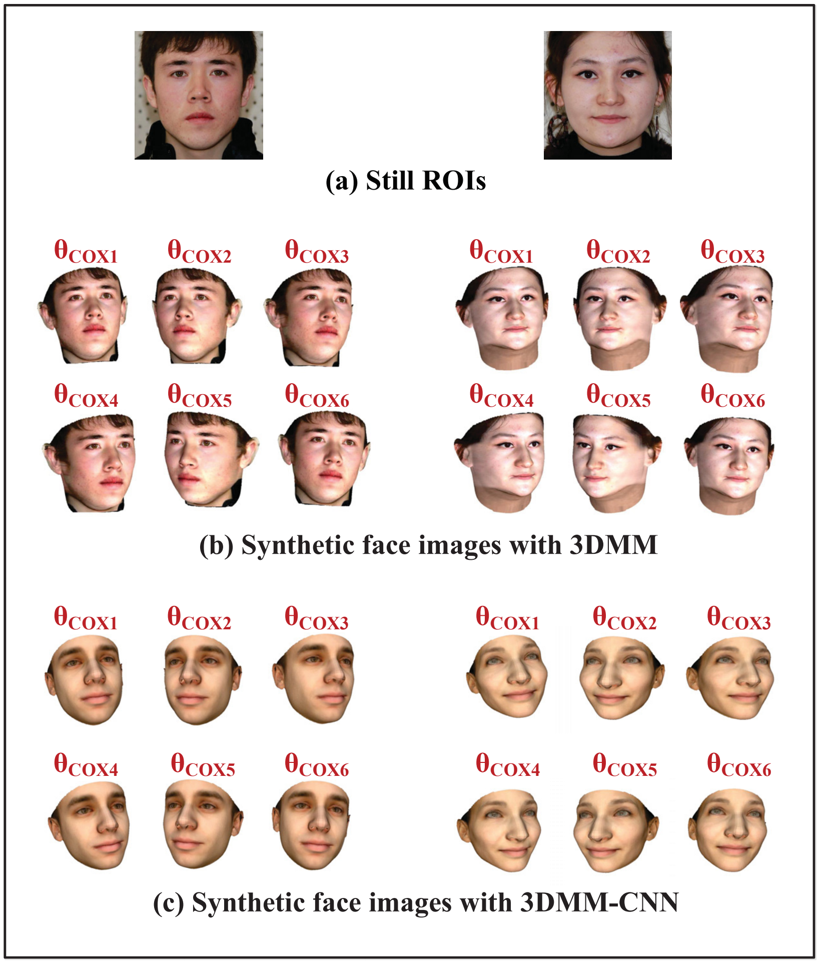
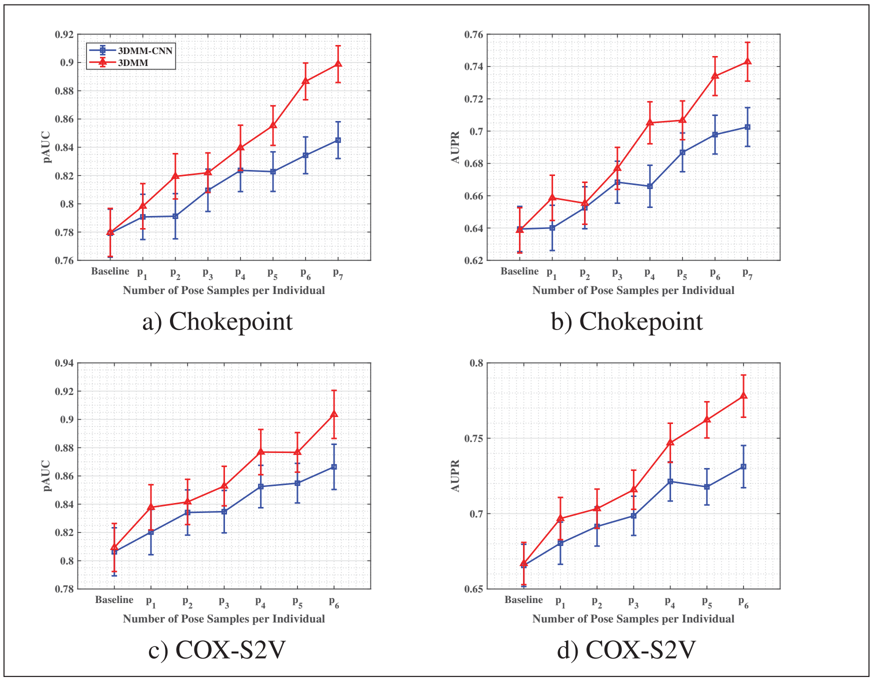

(a) Still ROIs
θCOX1
θCOX2
θCOX3
θCOX4
θCOX5
θCOX6
(b) Synthetic face images with 3DMM
θCOX1
θCOX2
θCOX3
θCOX4
θCOX5
θCOX6
(c) Synthetic face images with 3DMM-CNN

Figure 3.9 Examples of synthetic face images generated from the reference still ROI of individuals ID#21 and ID#151 (a) of COX-S2V dataset. They are produced based on representative exemplars (poses) and using 3DMM (b) and 3DMM-CNN (c).

siderably by about $20-30\%$ with only few synthetic ROIs (1 sample per pose cluster) for Chokepoint and COX-S2V datasets, respectively.

To further assess the benefits, Fig. 3.11 compares the performance of the proposed S+V method (adds q synthetic samples) with the original SRC (without an auxiliary dictionary), and to ESRC (with manually designed auxiliary dictionary). Results in this figure show that the proposed method outperforms the others, and that FR performance is higher when the dictionary is designed using the representative views than based on the manually designed dictionary. The



a) Chokepoint

b) Chokepoint

c) COX-S2V

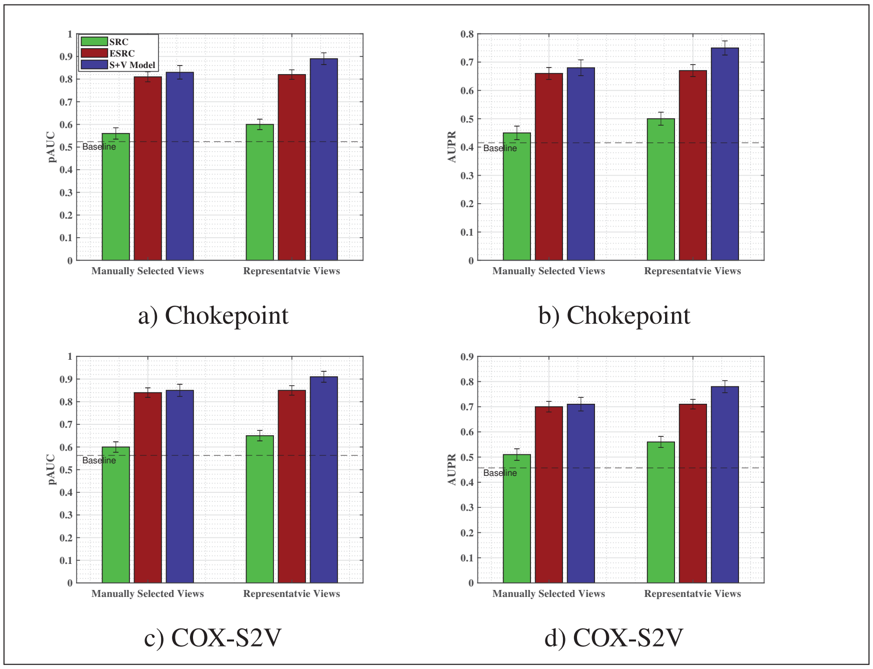
d) COX-S2V

Figure 3.10 Average pAUC(20%) and AUPR accuracy of S+V model versus the size of the synthetic set generated using 3DMM and 3DMM+CNN on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation.

proposed method can therefore adequately generate representative facial ROIs for the gallery, and then match it with the corresponding variations in the auxiliary dictionary. Encouraging pair-wise relationships between the variational and augmented gallery dictionaries has a positive impact on the performance of still-to-video FR system based on SRC.

3.6.3 Impact of Camera Viewpoint

To evaluate the robustness of the proposed S+V model to pose variations, accuracy is measured for different portals and video cameras, as well as for a fusion of cameras. Tables 3.1 and 3.2 summarize the average accuracy on Chokepoint and COX-S2V datasets, respectively. For the Chokepoint dataset, videos are captured over 4 sessions for 3 cameras (Camera1, Camera2, Camera3) over portals 1 (P1E, P1L) and portal 2 (P2E, P2L), while for the COX-S2V dataset, videos are captured over 3 cameras (Camera1, Camera2 and Camera3). The performance



a) Chokepoint

b) Chokepoint

c) COX-S2V

d) COX-S2V

Figure 3.11 Average pAUC(20%) and AUPR accuracy for SRC, ESRC and S+V model on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation.

of the S+V model is compared with that of SRC and ESRC using the same configurations. Results show that the S+V model outperforms other techniques across different pose variations. Using synthetic profile views can improve the robustness of FR systems to pose variations. As expected, designing a system that combines faces from all the cameras (and portals) always provides a higher level of accuracy.

3.6.4 Impact of Feature Representations

Table 3.3 shows the effect on FR performance of using different feature representations (including raw pixels, AlexNet (Krizhevsky *et al.* (2012)), ResNet (He *et al.* (2016)) and VGGNet (Simonyan & Zisserman (2015)) and face synthesis methods (3DMM and 3DMM-CNN) for videos from all 3 cameras of the Chokepoint and COX-S2V datasets.

Table 3.1 Average accuracy of FR systems based on the proposed S+V model, SRC, and ESRC over different sessions, portals and cameras of the Chokepoint dataset. Feature representations are raw pixels, the 3DMM method is used for face synthesis.

Portal	Viewpoint	Accuracy					
		SRC		**ESRC**		**S+V Model**	
		pAUC(20%)	AUPR	pAUC(20%)	AUPR	pAUC(20%)	AUPR
P1	Camera1	0.482±0.023	0.361±0.021	0.691±0.020	0.534±0.023	0.712±0.024	0.607±0.021
	Camera2	0.495±0.021	0.389±0.022	0.703±0.022	0.553±0.020	0.719±0.022	0.615±0.022
	Camera3	0.412±0.025	0.377±0.023	0.532±0.023	0.512±0.022	0.672±0.026	0.572±0.023
	All 3 Cameras	0.513±0.022	0.438±0.024	0.718±0.019	0.579±0.018	0.731±0.021	0.706±0.022
P2	Camera1	0.422±0.023	0.387±0.020	0.604±0.024	0.526±0.021	0.622±0.022	0.518±0.020
	Camera2	0.452±0.022	0.416±0.023	0.631±0.025	0.548±0.020	0.652±0.021	0.546±0.021
	Camera3	0.378±0.021	0.351±0.022	0.517±0.022	0.435±0.023	0.538±0.025	0.441±0.022
	All 3 Cameras	0.471±0.020	0.423±0.021	0.651±0.020	0.547±0.019	0.672±0.018	0.573±0.023
P1&P2	All 3 Cameras	0.524±0.032	0.475±0.031	0.802±0.028	0.651±0.025	0.892±0.019	0.751±0.020

Table 3.2 Average accuracy of FR systems using the proposed S+V model, SRC, and ESRC over different sessions and portals of the COX-S2V dataset. Feature representations are raw pixels, the 3DMM method is used for face synthesis.

Viewpoint	Accuracy					
	SRC		**ESRC**		**S+V Model**	
	pAUC(20%)	AUPR	pAUC(20%)	AUPR	pAUC(20%)	AUPR
Camera1	0.481±0.020	0.432±0.021	0.765±0.019	0.645±0.022	0.780±0.020	0.657±0.021
Camera2	0.475±0.023	0.419±0.022	0.716±0.020	0.602±0.020	0.747±0.023	0.629±0.022
Camera3	0.507±0.021	0.441±0.019	0.802±0.021	0.671±0.021	0.824±0.021	0.715±0.019
All 3 Cameras	0.566±0.030	0.480±0.027	0.835±0.027	0.695±0.026	0.905±0.020	0.776±0.017

We further evaluate the impact on the performance of different CNN feature extractors and loss functions for FR with the S+V model. Table 3.4 shows the average AUC and AUPR accuracy of FR systems using the proposed S+V model with different pre-trained CNNs for feature representation and loss functions (triplet loss (Schroff *et al.* (2015)), cosine loss (Liu *et al.* (2017)) and angular softmax (Wang *et al.* (2018a))) on the Chokepoint and COX-S2V databases. Results indicate that coupling the S+V model with deep CNN features can further improve FR accuracy over using raw pixels, and that using ResNet-50 outperforms there other CNN architectures. Additionally, SphereFace training method yields the higher accuracy. By using CNN features along with 3DMM or 3DMM-CNN, a still-to-video FR system with the S+V model outperforms the baseline template matcher (TM) and SRC.

Table 3.3 Average accuracy of FR systems using the proposed S+V model and template matching using different feature representation on Chokepoint and COX-S2V databases.

Technique	**Face Synthesis**	**Features**	**Accuracy**			
			Chokepoint database		**COX-S2V database**	
			pAUC(20%)	AUPR	pAUC(20%)	AUPR
TM	N/A	Raw pixels	0.551±0.027	0.503±0.028	0.574±0.031	0.512±0.029
		AlexNet	0.563±0.026	0.513±0.029	0.586±0.030	0.519±0.027
		VGGNet-16	0.570±0.028	0.524±0.026	0.597±0.027	0.528±0.030
		VGGNet-19	0.578±0.025	0.531±0.027	0.605±0.029	0.533±0.028
		ResNet-50	0.595±0.027	0.550±0.026	0.628±0.024	0.551±0.025
SRC	N/A	Raw pixels	0.525±0.030	0.475±0.029	0.568±0.031	0.481±0.030
		AlexNet	0.537±0.025	0.487±0.028	0.581±0.027	0.494±0.026
		VGGNet-16	0.552±0.026	0.491±0.027	0.590±0.025	0.505±0.027
		VGGNet-19	0.567±0.027	0.512±0.024	0.602±0.023	0.511±0.028
		ResNet-50	0.581±0.026	0.533±0.025	0.623±0.022	0.523±0.024
S+V Model	3DMM	Raw pixels	0.892±0.018	0.751±0.019	0.903±0.020	0.775±0.016
		AlexNet	0.905±0.019	0.771±0.020	0.913±0.016	0.783±0.015
		VGGNet-16	0.908±0.016	0.773±0.017	0.916±0.018	0.786±0.016
		VGGNet-19	0.912±0.017	0.779±0.018	0.921±0.016	0.791±0.017
		ResNet-50	**0.917±0.015**	**0.783±0.016**	**0.925±0.015**	**0.798±0.014**
	3DMM-CNN	Raw pixels	0.855±0.019	0.737±0.018	0.871±0.019	0.741±0.018
		AlexNet	0.873±0.020	0.752±0.020	0.884±0.018	0.753±0.019
		VGGNet-16	0.880±0.017	0.759±0.017	0.891±0.017	0.761±0.016
		VGGNet-19	0.884±0.018	0.763±0.020	0.902±0.016	0.765±0.017
		ResNet-50	0.891±0.016	0.769±0.014	0.907±0.017	0.771±0.015

Results show that coupling the S+V model with deep CNN features can further improve the FR accuracy over using raw pixels, and that using ResNet-50 outperforms all other deep architectures. The results also indicate that SphereFace training method yields higher accuracy. Using CNN features and 3DMM or 3DMM-CNN, a FR system with the S+V model outperform the baseline template matcher (TM) and SRC.

Tables 3.5 shows the average accuracy of FR for the augmented and auxiliary dictionaries with the videos from all 3 cameras of the Chokepoint and COX-S2V datasets, respectively.

3.6.5 Comparison with State-of-the-Art Methods

Table 3.6 presents the FR accuracy obtained with the proposed S+V model compared with baseline SRC techniques based on generic learning – ESRC (Deng *et al.* (2012)), SVDL (Yang *et al.* (2013)), LGR (Zhu *et al.* (2014)), RADL (Wei & Wang (2015)), CSR (Li *et al.* (2016)).

Table 3.4 Average accuracy of FR systems using the proposed S+V model (3DMM face synthesis) with different deep feature representations on Chokepoint and COX-S2V databases.

Technique	**Deep Architecture**	**Training**	**Accuracy**			
			Chokepoint database		**COX-S2V database**	
			pAUC(20%)	AUPR	pAUC(20%)	AUPR
S+V Model	AlexNet	FaceNet	0.905±0.019	0.771±0.020	0.913±0.016	0.783±0.015
		CosFace	0.908±0.021	0.774±0.022	0.915±0.017	0.787±0.016
		SphereFace	0.912±0.020	0.780±0.018	0.918±0.015	0.792±0.014
	VGGNet-19	FaceNet	0.884±0.021	0.763±0.020	0.902±0.019	0.765±0.018
		CosFace	0.889±0.019	0.768±0.022	0.907±0.017	0.772±0.016
		SphereFace	0.906±0.018	0.771±0.017	0.913±0.015	0.778±0.017
	ResNet-50	FaceNet	0.917±0.015	0.783±0.016	0.924±0.015	0.798±0.014
		CosFace	0.920±0.018	0.786±0.019	0.927±0.018	0.802±0.020
		SphereFace	0.922±0.015	0.791±0.014	0.928±0.017	0.805±0.015

Table 3.5 Average accuracy of FR systems using the augmented dictionary (3DMM face synthesis) and auxiliary dictionaries on Chokepoint and COX-S2V databases.

Technique		**Accuracy**			
		Chokepoint database		**COX-S2V database**	
		pAUC(20%)	AUPR	pAUC(20%)	AUPR
S+V Model	Augmented Dictionary	0.829±0.28	0.705±0.27	0.847±0.26	0.718±0.254
	Auxiliary Dictionary	0.836±0.23	0.714±0.25	0.862±0.22	0.731±0.021

Each one uses the same number of samples, raw pixel-based features, and a regularization parameter λ set to 0.005. Accuracy of the S+V model is also compared with that of the Flow-Based Face Frontalization (Hassner *et al.* (2015)) and Recognition via Generation (Masi *et al.* (2016)) techniques. The baseline system is a SRC model designed with the original reference still ROI of each enrolled person, and raw pixel-based features. The table shows that the S+V model, using a joint generic learning and face synthesis, achieves the higher level of accuracy than other methods under the same configuration, has potential in surveillance FR.

In order to assess still-to-video FR accuracy under the worst-case pose variations between the probe video ROIs and augmented gallery dictionary ROIs, we compute the minimum distance between the pose angle of each probe video ROI (20 trajectories in 3 cameras), $\{\theta_1, \theta_2, \ldots, \theta_n\}$, and pose angles of both reference still and synthetic ROIs in the augmented gallery dictionary,

Table 3.6 Average accuracy of FR systems based on the proposed S+V model and related state-of-the art SRC methods for videos from all 3 cameras of the Chokepoint and COX-S2V databases. Feature representations are raw pixels, the 3DMM method is used for face synthesis.

Techniques	**Accuracy**			
	Chokepoint database		**COX-S2V database**	
	pAUC(20%)	AUPR	pAUC(20%)	AUPR
SRC (Wright *et al.* (2009))	0.524±0.032	0.475±0.031	0.568±0.030	0.480±0.027
ESRC (Deng *et al.* (2012))	0.802±0.028	0.651±0.025	0.835±0.027	0.695±0.026
ESRC-KSVD	0.811±0.023	0.659±0.022	0.840±0.023	0.712±0.021
SVDL (Yang *et al.* (2013))	0.825±0.023	0.703±0.025	0.843±0.025	0.724±0.023
RADL (Wei & Wang (2015))	0.832±0.019	0.711±0.020	0.849±0.022	0.730±0.021
LGR (Zhu *et al.* (2014))	0.849±0.022	0.717±0.024	0.878±0.023	0.744±0.025
CSR (Li *et al.* (2016))	0.852±0.025	0.722±0.020	0.880±0.021	0.753±0.020
Face Frontalization (Hassner *et al.* (2015))	0.822±0.021	0.711±0.023	0.843±0.022	0.719±0.023
Recognition via Generation (Masi *et al.* (2016))	0.815±0.023	0.703±0.025	0.838±0.024	0.705±0.026
S+V Model (Ours)	**0.892±0.019**	**0.751±0.020**	**0.905±0.018**	**0.776±0.017**

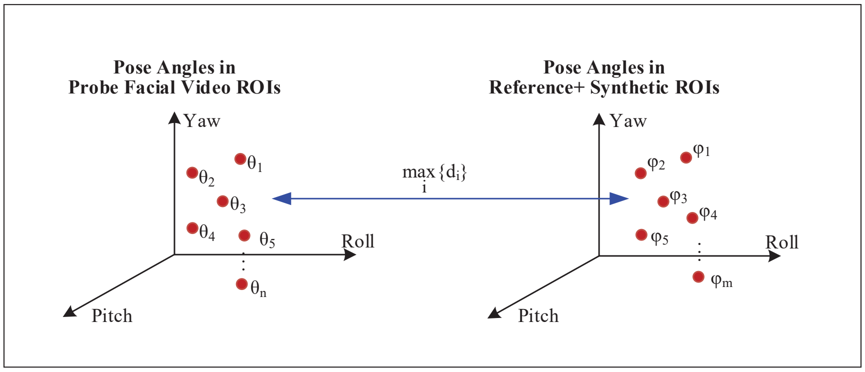


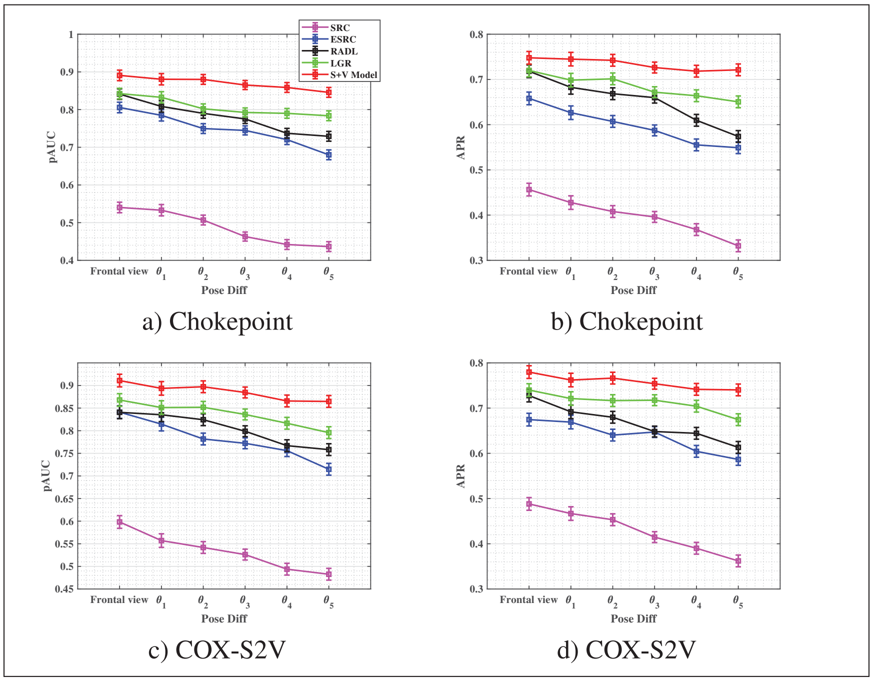
Figure 3.12 Illustration of procedure for the selection of the largest pose variations.

$\{\varphi_1, \varphi_2, \ldots, \varphi_m\}$:

$$d_i = \min_j \{\| (\theta_i - \varphi_j) \| : j = 1,2,\ldots,m\}, \; . \tag{3.20}$$

where d_i corresponds to the i^{th} probe video ROI, for $i = 1,2,\ldots,n$. Next, 5 video ROIs that have the largest distance, $\max_i\{d_i\}$, are chosen as the faces with the largest pose differences (see Fig. 3.11). Fig. 3.13 shows the accuracy obtained with the SRC, ESRC, RADL, LGR and S+V models when these ROIs are classified as probe ROIs.

As the pose differences increase, FR accuracy decreases. The FR system using the S+V model reaches the highest accuracy due to the added robustness to pose variations. Then, LGR outper-



a) Chokepoint

b) Chokepoint

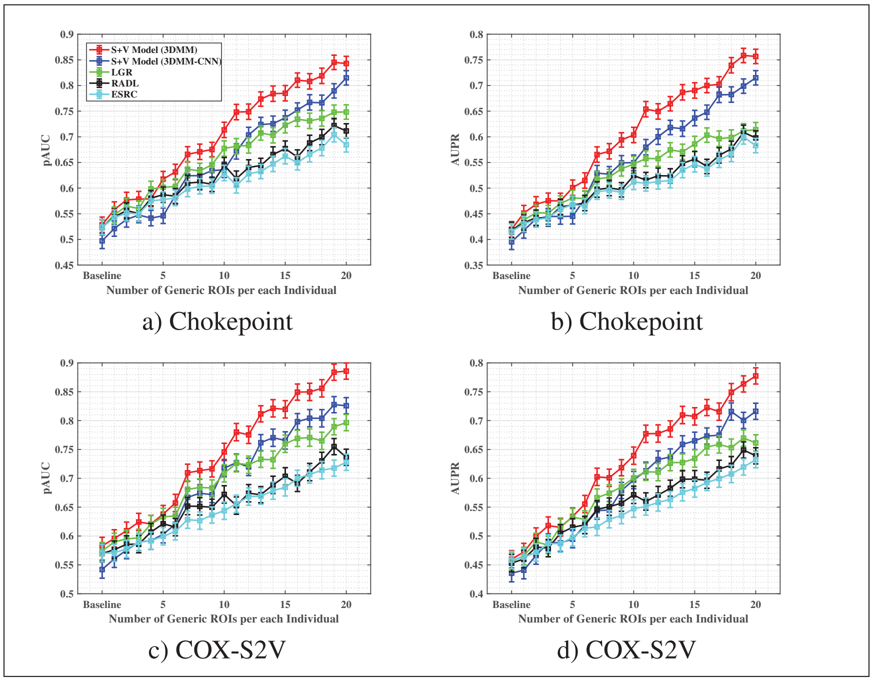
c) COX-S2V

d) COX-S2V

Figure 3.13 Average pAUC(20%) and AUPR accuracy of S+V model and related state-of-the-art techniques versus the different pose variations on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation.

forms SRC, ESRC and RADL across all pose variations. Accuracy of the SRC is much lower than the others because, with only one frontal reference gallery ROI per person, the probe ROIs are not well represented.

Fig. 3.14 shows the impact of the size of generic set in the auxiliary variational dictionary on FR accuracy. The results of SRC, ESRC, RADL and LGR are also shown for the same configurations for comparison. Accuracy of the S+V model increases significantly with respect to other state-of-the-art methods as the number of generic ROIs grows. The results support the conclusion that by augmenting the gallery dictionary, allows the S+V model to increasingly benefit from the variational information of the generic set.



a) Chokepoint

b) Chokepoint

c) COX-S2V

d) COX-S2V

Figure 3.14 Average pAUC(20%) and AUPR accuracy versus the size of the generic set on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation.

3.6.6 Ablation Study

Designing S+V model for still-to-video FR consists of three main steps: ($\mathcal{M}_1$) face synthesis, ($\mathcal{M}_2$) adding intra-class variations, and ($\mathcal{M}_3$) pairing the dictionaries. In this subsection, an ablation study is presented to show the impact of each module on the FR performance. We assume that all FR systems use a pixel-based feature representation, 3DMM face synthesis, and q synthetic images in the augmented dictionary.

Tables 3.7 and 3.8 shows the average accuracy of the ablation study with videos from all 3 cameras of the Chokepoint and COX-S2V datasets, respectively. Firstly, we disabled the face synthesis module, $\mathcal{M}_1$, and performed experiments to show the impact of augmenting the reference gallery with synthetic faces on FR accuracy. Next, we removed the auxiliary dictionary to evaluate the impact of considering generic set variations with the S+V model. By removing both $\mathcal{M}_1$ and $\mathcal{M}_2$ modules from the S+V model, accuracy declines significantly by about 50%.

The results suggest that the addition of synthetic and generic set faces is an effective strategy to cope with facial variations. Another important component of the S+V model is the selection of representative ROIs and pairing the dictionaries. By removing the row sparsity and joint sparsity in the S+V model, $\mathcal{M}_3$, and by adding 10 randomly selected synthetic ROIs, accuracy decreases by about 15%.

Table 3.7 Results of ablation study with Chokepoint database.

Accuracy	**Removed Module**			
	baseline (none)	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
pAUC(20%)	0.892±0.019	0.839±0.21	0.827±0.27	0.883±0.25
AUPR	0.751±0.020	0.709±0.23	0.702±0.25	0.721±0.22

Table 3.8 Results of ablation study with COX-S2V database.

Accuracy	**Removed Module**			
	baseline (none)	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
pAUC(20%)	0.905±0.018	0.857±0.22	0.835±0.24	0.887±0.20
AUPR	0.776±0.017	0.721±0.20	0.712±0.21	0.769±0.21

3.6.7 Complexity Analysis

Time complexity is an important consideration in many real-time FR applications in video surveillance. The time required by the S+V model to classify a probe ROI is $\mathcal{O}(d(N+M)Lq\log n + Lk(q+1))$ where d is the dimension of the face descriptors, n is the number of ROIs per class in the augmented gallery dictionary, k is the number of classes (enrolled individuals), $N = kn$ is the number of reference still images, M is the size of the external generic set, q is the number of views, and L is number active sets (at each iteration, we need to select L most representative dynamic active sets from coefficient matrix.) In video FR applications, N may be larger, therefore the computational burden of handling larger dictionaries may represent bottleneck of the proposed method. The complexity of SRC and ESRC are $\mathcal{O}(d^2N)$, $\mathcal{O}(d^2(N+M))$, respectively. The complexity of LGR is $\mathcal{O}(s({n_d}^3 + {n_d}^2 d_p))$ where s is the number of patches, n_d is the number of patches, d_p is the feature dimension of patches. Although the proposed S+V model outperforms SRC and ESRC, it requires more computations, mostly because of the pairing of the dictionaries.

Table 3.9 reports the average test time required by the proposed and baseline techniques to classify a probe ROI from Chokepoint and COX-S2V videos. The LGR and RADL are more computationally intensive than the S+V model. Finally, Table 3.10 reports the average time for the 3 main steps of the proposed framework: face synthesis ($\mathcal{M}_1$), intra-class variation extraction ($\mathcal{M}_2$), and pairing the dictionaries ($\mathcal{M}_3$) on videos of all 3 cameras in the Chokepoint and COX-S2V datasets. The time complexity of $\mathcal{M}_1$ is the highest, followed by $\mathcal{M}_3$ with complexity $\mathcal{O}(MNlog(M))$, where M and N are, respectively, the number of rows and columns of the dissimilarity matrix.

Table 3.9 Average time required by techniques to classify a probe videos ROI with the Chokepoint and COX-S2V datasets.

Technique	**Classification Time (sec)**	
	Chokepoint database	**COX-S2V database**
SRC (Wright *et al.* (2009))	1.03	2.56
ESRC (Deng *et al.* (2012))	1.72	3.42
RADL (Wei & Wang (2015))	4.62	8.15
LGR (Zhu *et al.* (2014))	7.13	12.37
S+V Model	2.81	4.83

Table 3.10 Average computational time of different step in the S+V model with the Chokepoint and COX-S2V datasets.

Module	**Processing Time (Sec)**	
	Chokepoint database	**COX-S2V database**
$\mathcal{M}_1$ (3DMM)	120	120
$\mathcal{M}_1$ (3DMM-CNN)	1.3	1.3
$\mathcal{M}_2$	0.53	0.53
$\mathcal{M}_3$	2.47	4.41

3.7 Conclusion

In this paper, a paired sparse reconstruction model is proposed to account for linear and non-linear variations in the context of still-to-video FR. The proposed S+V model leverages both face synthesis and generic learning to effectively represent probe ROIs from a single reference still. This approach manages the non-linear variations by enriching the gallery dictionary with a representative set of synthetic profile faces, where synthetic (still) faces are paired with generic set (video) face in the auxiliary variational dictionary. In this way, the augmented gallery dictionary is encouraged to share the same sparsity pattern with the auxiliary dictionary for the same pose angles. Experimental results obtained using the Chokepoint and COX-S2V datasets suggest that the proposed S+V model allows us to efficiently represent linear and non-linear variations in facial pose with no need to collect a large amount of training data. Results indicated that generic learning alone cannot effectively resolve the challenges of the SSPP and visual domain shift problems. With S+V model, generic learning and face synthesis are complementary. The results also reveal that the performance of FR systems based on the S+V model can further improve with CNN features. Future research includes investigating the geometrical structure of the data space in the dictionaries and the corresponding coefficients to improve the discrimination. To reduce reconstruction time, we plan to extend the current S+V model, allowing it to represent larger sparse codes.

CHAPTER 4

VIDEO FACE RECOGNITION USING SIAMESE NETWORKS WITH BLOCK-SPARSITY MATCHING

Fania Mokhayeri, Eric Granger

Le Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA),
École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article published in « IEEE Transactions on Biometrics, Behavior, and Identity Science »
2019.

Abstract

Deep learning models for still-to-video FR typically provide a low level of accuracy because faces captured in unconstrained videos are matched against a reference gallery comprised of a single facial still per individual. For improved robustness to intra-class variations, deep Siamese networks have recently been used for pair-wise face matching. Although these networks can improve state-of-the-art accuracy, the absence of prior knowledge from the target domain means that many images must be collected to account for all possible capture conditions, which is not practical for many real-world surveillance applications. In this paper, we propose the deep SiamSRC network that employs block-sparsity for face matching, while the reference gallery is augmented with a compact set of domain-specific facial images. Prior to deployment, clustering based on row sparsity is performed on unlabelled faces captured in videos from the target domain. Cluster centers discovered in the capture condition space (defined by, e.g., pose, scale and illumination) are used as rendering parameters with an off-the-shelf 3D face model, and a compact set of synthetic faces are thereby generated for each reference still based on representative intra-class information from the target domain. For pair-wise similarity matching with query facial images, the SiamSRC exploits sparse representation-based classification with a block structure. Experimental results obtained with the videos from the Chokepoint and COX-S2V datasets indicate that the proposed SiamSRC network can outper-

form state-of-the-art methods for still-to-video FR with a single sample per person, with only a moderate increase in computational complexity.

4.1 Introduction

Video face recognition (FR) has attracted much interest from academia and industry due to the wide range of monitoring, security and surveillance applications. In this paper, we focus on deep learning architectures for still-to-video FR, where systems seek to match each query ROI captured in unconstrained videos (from the target domain) to still facial ROIs stored in a reference gallery (from the source domain) (Bashbaghi *et al.* (2017a)). Architectures based on the convolutional neural network (CNN) have achieved state-of-the-art performance across a wide range of visual recognition tasks like FR, at the expense of growing network complexity (Cao *et al.* (2019a)). Training CNNs with large datasets of facial images allow to encode discriminant feature embeddings that can provide accurate predictions during inference (Wang *et al.* (2018a)).

While deep learning architectures can provide a high level of accuracy, designing robust systems for still-to-video FR remains a challenging problem in real-world video surveillance applications, such as watch-list screening (Dewan *et al.* (2016)). One key challenge is the limited number of reference still images that are available per individual enrolled to the system, and stored in a gallery. Still-to-video FR systems are typically designed using only one still image per individual. This corresponds to a single sample per person (SSPP) problem in FR, where facial models have limited robustness to intra-class variations. State-of-the-art approaches proposed to address SSPP problems in FR systems can be roughly divided into four categories: (1) image patching methods, where the images are partitioned into several patches (Zhu *et al.* (2014); Gao *et al.* (2015)), (2) multiple face representations that extract diverse features from face images where each descriptor may be specialized to address some nuisance factors (e.g., illumination, pose, blur, etc.) encountered in video surveillance (Bashbaghi *et al.* (2017b)), (3) super resolution (hallucination) that aims at reconstructing details/high-frequencies in low-resolution face images (Yu *et al.* (2018)), (4) face synthesis techniques to augment the gallery dictionary (Mokhayeri *et al.* (2019a); Masi *et al.* (2016)), and (5) generic learning methods, where a genetic training set is used to leverage variational information from an auxiliary generic set to represent the differences between probe and reference gallery images (Wei & Wang (2015); Deng *et al.* (2018)).

Another key challenge for still-to-video FR is the visual domain shift in the distributions between facial ROIs from the source domain, where reference stills are typically captured during enrolment under controlled conditions, and facial ROIs from the target domain, where videos are captured under uncontrolled conditions with variations in pose, illumination, motion blur, scale, resolution, expression, etc. The appearance of facial ROIs captured in the videos correspond to target domain distribution can differ considerably from reference faces captured during enrollment in a gallery (Bashbaghi *et al.* (2017a)). A network for still-to-video FR would require transfer learning or unsupervised domain adaptation to learn robust domain-invariant representations from source and target ROIs. Recent state-of-the-art approaches for still-to-video FR rely on deep Siamese networks (Koch *et al.* (2015)) for pair-wise matching. These architectures apply two identical CNNs (same parameters and weights) to extract features from the reference gallery and query facial ROIs, and then compare the feature vectors using some similarity measures. They are trained to learn embeddings where similar image pairs (portraying the same identity) are close to each other, while dissimilar image pairs (with different identities) are distant from each other (Ahmed *et al.* (2015); Varior *et al.*; Parchami *et al.* (2017c)). For instance, DeepFace (Taigman *et al.* (2014)) employs a deep Siamese network to compare the feature vectors of the gallery and query faces using the Euclidean distance. The goal of training is to minimize the distance between congruous pairs of faces (i.e. portraying the same identity) and maximize the distance between incongruous pairs. However, achieving state-of-the-art performance comes with high computational complexity. For instance, the ResNet-50 CNN (He *et al.* (2016)) contains about 23.5M parameters (stored in about 86MB of memory), and requires 6.3 billion floating point operations (FLOPs) to match two color images of size $256 \times 128 \times 3$. Such complex networks are impractical for many real-time applications. Consequently, even if all reference stills are stored in the gallery as feature vectors, there is an interest in reducing the number of pair-wise matchings needed by the network. The time and memory complexity of a Siamese network grows with the number of reference gallery images, and thus of pair-wise matchings.

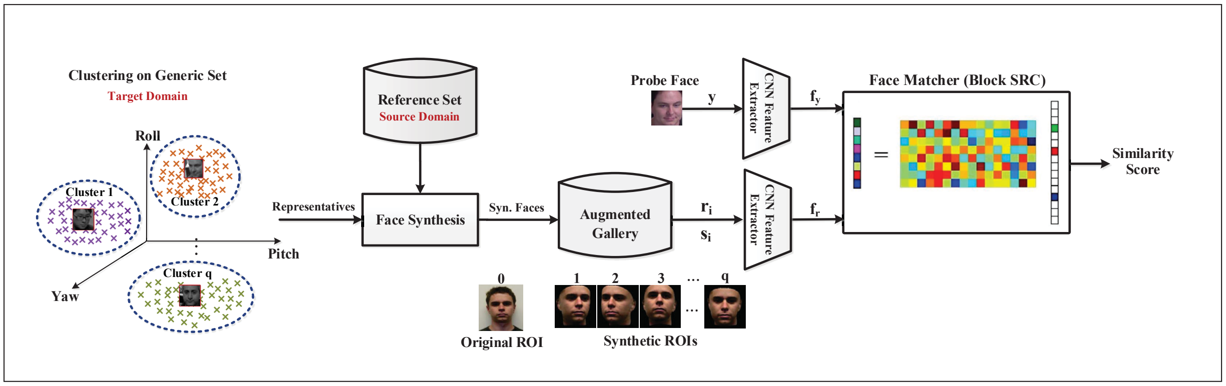


Figure 4.1 An overview of the proposed deep SiamSRC network for still-to-video FR.

In this paper, deep SiamSRC network is proposed (see Figure 4.1) that uses sparse representation-based classification (SRC) for pair-wise face matching while integrates domain-specific set of synthetic facial images into the network's extended reference gallery. The motivation is to improve the FR system's robustness to intra-class variations of individual appearing in the target domain and compensate limited images available in the gallery. One concern with data augmentation is the selection of a sufficient number of synthetic or generic auxiliary faces to cover intra-class variations. Many synthetic (generic) faces may be generated (collected) to account for all possible real-world capture conditions, although several of these may provide less relevant information on the target domain. FR systems would therefore require high computational complexity to accommodate many reference gallery images. To provide a good trade-off between accuracy and efficiency, we select representative synthetic ROIs by exploiting the discriminant information of the generic set during the face synthesis process. Prior to deployment (e.g., during camera calibration), the generic set is formed with facial ROIs captured in target domain videos, and clustered based on row sparsity (Elhamifar *et al.* (2012)) in a captured condition space. The gallery is augmented with a set of synthetic face images generated from the original reference image in the source domain, where the rendering parameters are estimated based on the cluster centers (exemplars) in the target video domain. During inference, the proposed SiamSRC network exploits the SRC for pair-wise face matching. SRC similarity is measured as the block structure that finds the representation of a query ROI and requires the minimum number of blocks from the gallery.

For proof-of-concept validation, the proposed SiamSRC network relies on representative synthetic sets that are selected by clustering the generic set in the pose and illumination space. Performance is evaluated and compared to several relevant state-of-the-art methods on two public databases for still-to-video FR – Chokepoint (Wong *et al.* (2011)) and COX-S2V (Huang *et al.* (2015)).

The rest of the paper is organized as follows. Section 4.2 provides a brief review of deep Siamese networks, SRC models and face synthesizing methods that address still-to-video FR problems. Section 4.3 describes the proposed deep SiamSRC network. Finally, Sections 4.4 and 4.5 describe the experimental methodology and results, respectively.

4.2 Related Work

4.2.1 Deep Siamese Networks for Face Recognition

The idea of using deep Siamese networks for pair-wise matching of query and reference images in biometric authentication and verification originates from Bromley *et al.* (1994). Deep Siamese networks are often designed using two or more identical sub-networks for feature extraction. These extractors share the same parameters and weights, and are commonly implement with CNNs suitable for classification. During the training process, these networks typically seek to minimize the intra-class distance and maximize the inter-class distances mostly using triplet loss function (Schroff *et al.* (2015)). When the features are extracted for a pair of images, the matcher produces a similarity score indicating if the pair images are from the same or different classes. This similarity measure is often the Euclidean or cosine distance between the two feature representations (Wang *et al.* (2018a)). Similarity can also be assessed using a Softmax layer with two classes, where one neuron represents the same class and the other different class. For a given query image, the system provides a ranked list of matching scores for every reference image in the gallery, where the highest score represents the network's prediction for the input image. The feature extractor sub-network is considered to be the backbone of a Siamese network. Various CNN architectures have been proposed to learn discriminative feature embeddings, most of which employ end-to-end training, where both feature embeddings and metrics are learned as a joint optimization problem (Ahmed *et al.* (2015); Varior *et al.*). Any CNN such as VGG-Face, Inception, ResNet, DenseNet, etc., can be used for feature extraction. Although deeper CNNs, like ResNet-50 and DenseNet, are better to address the challenges of real-world FR, the computational complexity and over-fitting make them less suitable for real-time applications.

In recent years, deep Siamese networks have significantly improved FR accuracy due to their high capacity for learning discriminative features. Taigman *et al.* proposed to use these networks to learn similarity metrics for FR which trained on a large dataset from Facebook (Taigman *et al.* (2014)). Schroff *et al.* introduced FaceNet that directly learns a mapping from face images to a compact Euclidean space. It employs a deep Siamese network that directly optimizes the L_1-distance between two faces. FaceNet employs face triplets and minimizes the distance between an anchor and a positive sample of the same identity, while maximizing the distance between the anchor and a negative sample of a different identity (Schroff *et al.* (2015)). Light CNN framework was proposed to learn deep face representations from the large-scale dataset with noisy labels, where a max-feature map operation allows to obtain a compact representation (Wu *et al.* (2018a)). Yin & Liu (2017) presented a multi-task CNN for FR that exploits side tasks, e.g., pose, in regularization, to learn pose-specific identity features. A pose-aware network was proposed by Masi *et al.* (2019a) to process a face image using several pose-specific CNNs. In this model, 3D rendering was used to synthesize multiple face poses from input images to train these models, and provide additional robustness to pose variations. A deep local descriptor learning framework for cross-modality FR was presented by Peng *et al.* (2019) to learn discriminant and compact local information from raw facial patches, and then integrated into a CNN that extracts deep local descriptors. Peng *et al.* (2019) proposed an efficient network for still-to-video FR from a single reference still based on cross-correlation matching and triplet-loss optimization methods that provide discriminant face representations. The matching pipeline exploits a matrix Hadamard product followed by a fully connected layer inspired by adaptive weighted cross-correlation.

4.2.2 Face Synthesis

Augmenting the reference gallery set synthetically is known to improve the accuracy of CNN-based methods. Masi *et al.* (2019b) enhanced CNN performance by augmenting dataset with facial images that differ in 3D shape, expression and pose. An efficient face-specific data augmentation technique has been introduced by Masi *et al.* (2019b) to enrich the training data for CNN-based FR to reduce appearance variations. A common approach for synthetic face generation is to reconstruct the 3D model of a face using its 2D face image. Blanz & Vetter (2003) proposed 3D Morphable Model (3DMM) to reconstruct a 3D face from a single 2D face image and accordingly synthesize new face images. A CNN was proposed by Tran *et al.* (2017a) to regress 3DMM shape and texture parameters directly from an input image without an optimization process which renders the face, and compares it to the image. Richardson *et al.* (2017) presented a face reconstruction technique from a single image by introducing an end-to-end CNN framework which derives the shape in a coarse-to-fine fashion. Tewari *et al.* (2017) proposed a model-based deep convolutional autoencoder for 3D face reconstructing from a single, where a convolutional encoder network is combined with an expert-designed generative model that serves as a decoder. Mokhayeri *et al.* (2019a) proposed a domain-specific face synthesis technique that exploits the representative intra-class variation information available from the target domain. It maps the intra-class variations from a representative set of video ROIs selected from the target domain into the original reference still ROIs.

Recently, Generative Adversarial Networks (GANs) proposed by Goodfellow & et al. (2014) have shown promising performance in face synthesis. Benefiting from GAN, FaceID-GAN (Shen *et al.* (2018)) was proposed which generates identity preserving faces. It competes with the generator by distinguishing the identities of the real and synthesized faces to preserve the identity of original images. The major shortcoming of the GAN-based face synthesis models is that they may produce images that are inconsistent due to the weak global constraints. To reduce this gap, Shrivastava *et al.* (2017) developed SimGAN that learns a model using synthetic images as inputs instead of random noise vector. Recently, conditional GANs allow to leverage conditional information in the generative and discriminative networks for conditional image synthesis. Tran *et al.* (2018) used pose codes in conjunction with random noise vectors as the inputs to the discriminator with the goal of generating a face of the same identity with the target pose in order to fool the discriminator. Hu *et al.* (2018) introduced a coupled-agent discriminator which forms a mask image to guide the generator during the learning process. Chen & Ross (2019) proposed semantic-guided generative adversarial network to automatically synthesize face images. In particular, semantic labels, extracted by a face parsing network, are used to compute a semantic loss function to regularize the adversarial network during training.

4.2.3 Sparse Representation-based Classification

Shao *et al.* (2017) presented a SRC-based FR algorithm that extends the dictionary using a set of synthetic faces generated by calculating the image difference of a pair of facial images. Deng *et al.* (2012) introduced an extended SRC (ESRC) that integrates an auxiliary variational dictionary (through random selection from a generic set) to accurately represent a probe face with unknown variations from the operational environment. Motivated by ESRC, Yang *et al.* (2013) proposed the sparse variation dictionary learning (SVDL) model to learn the variational dictionary by accounting for the relationship between the reference gallery and external generic set. A robust auxiliary dictionary learning (RADL) technique was proposed by Wei & Wang (2015) that extracts representative information from external data via dictionary learning without assuming the prior knowledge of occlusion in probe images. A collaborative probabilistic generic learning technique was introduced by Ji *et al.* (2017) which constructs probabilistic labels for the samples in the generic set corresponding to those in the reference gallery set, then it estimates the variation type for a given probe image. Deng *et al.* (2018)

developed a superposed linear representation classifier, where the test images are represented in terms of a superposition of the class centroids and the shared intra-class differences. A local generic representation-based (LGR) framework for FR with SSPP was proposed by Zhu *et al.* (2014). It builds a gallery dictionary by extracting the patches from the gallery database, while an intra-class variation dictionary is formed by using an external generic set to predict the possible facial variations. A joint and collaborative sparse representation framework is presented by Yang *et al.* (2017) that exploits the distinctiveness and commonality of different local regions. In order to address non-linearity, Fan *et al.* (2018) used a nonlinear mapping to transform the original reference data into a high dimensional feature space, which is achieved using a kernel-based method. Mokhayeri & Granger (2018) proposed a synthetic plus variational (S+V) model to account for the non-linearities. It reconstructs a probe image using an auxiliary variational dictionary and an augmented gallery dictionary while the dictionaries are encouraged to share the same sparsity pattern for the same pose angles.

4.3 The SiamSRC Network

This section describes the deep SiamSRC network for still-to-video FR. It relies on block sparsity to measure pair-wise similarity, and on a reference gallery that is augmented with a compact domain-specific set of synthetic images in order to overcome the issues related to visual domain shift and SSPP.

We select video ROIs with representative capture conditions (defined by pose angles and illumination measures) from facial trajectories or tracklets of unknown persons captured in the target domain. These video ROIs are selected by clustering facial trajectories with row sparsity in the captured condition space. A set of synthetic face images is generated according to cluster centers (representatives) to augment the reference gallery. A deep Siamese network is designed with the augmented gallery, where SRC with block structure is used for pair-wise face matching. The rest of this section presents more details on the proposed network.

4.3.1 Notation

In the following, the set $\mathbf{D} = \{\mathbf{R}_1, \ldots, \mathbf{R}_i, \ldots, \mathbf{R}_k\} \in \mathbb{R}^{d \times N}$ denote a gallery dictionary, where $\mathbf{R}_i = \{\mathbf{r}^i_1, \ldots, \mathbf{r}^i_j, \ldots, \mathbf{r}^i_n\} \in \mathbb{R}^{d \times n}$ is composed of n reference still ROIs belonging to one of k different classes, d is the number of pixels or features representing a ROI and $N = kn$ is the total number of reference still ROIs. In the context of SSPP problems $\mathbf{R}_i = \{\mathbf{r}^i\} \in \mathbb{R}^d$ is the single gallery sample of i^{th} class. The set $\mathbf{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_j \ldots, \mathbf{g}_M\} \in \mathbb{R}^{d \times M}$ denotes the auxiliary generic set composed of M external generic images of unknown persons captured in the operational environment. The set $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_j, \ldots, \mathbf{v}_M\} \in \mathbb{R}^{d \times M}$ denotes the auxiliary variational dictionary composed of M intra-class variations extracted from $\mathbf{G} \in \mathbb{R}^{d \times M}$.

4.3.2 Representative Selection

Prior to operation, e.g., during a camera calibration process, facial ROIs are isolated in facial trajectories from the videos of unknown persons captured in the target domain. A representative set of video ROIs are selected by applying row sparsity regularized optimization on facial trajectories in the captured condition space. A compact set of synthetic images is then generated from the reference set in the source domain based on the information obtained from the center of each cluster in the target domain, and integrated into the gallery dictionary to enrich the diversity of the gallery set. The representative selection problem is formulated as a row sparsity regularized trace minimization problem where the objective is to find a few representatives (exemplars) that efficiently represent the collection of data points according to their dissimilarities (Elhamifar & Kaluza (2017)). In this paper, the proposed model selects illumination and pose representatives from a collection of N samples. The pose angles are estimated using the discriminative response map fitting method (Asthana *et al.* (2013)) which is a state-of-the-art method for accurate fitting, suitable for handling occlusions and changing illumination conditions. The estimated head pose for the j^{th} video ROI ($\mathbf{g}_j$) in the generic set is defined as $\boldsymbol{\theta}_j = (\theta^{pitch}_j, \theta^{yaw}_j, \theta^{roll}_j)$. Euler angles θ^{pitch}, θ^{yaw}, and θ^{roll} are used to represent roll, yaw and pitch rotation around X axis, Y axis, and Z axis of the global coordinate system, respectively. Luminance and contrast distortion measures are also estimated between a video

ROI and the corresponding reference still ROI. Structural similarity index measure presented by Wang *et al.* (2004) are employed to measure the proximity of the average luminance and contrast locally by utilizing sliding window. The set of dissimilarities $\{d_{ij} : i, j = 1, ..., N\}$ between every pair of pose and illumination data points are then calculated by using the Euclidean distance, which indicates how well the data point i is suited to be an exemplar of data point j. The dissimilarities are arrange into matrix:

$$\mathbf{D} \triangleq \begin{bmatrix} \mathbf{d}_1^T \\ \vdots \\ \mathbf{d}_N^T \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{bmatrix} \in \mathbb{R}^{N \times N}, \tag{4.1}$$

where $\mathbf{d}_i \in \mathbb{D}^N$ denotes the i^{th} row of $\mathbf{D}$. Variables z_{ij} are associated with dissimilarities d_{ij}, and organized into matrix of the same size as:

$$\mathbf{Z} \triangleq \begin{bmatrix} \mathbf{z}_1^T \\ \vdots \\ \mathbf{z}_N^T \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{NN} \end{bmatrix} \in \mathbb{R}^{N \times N}, \tag{4.2}$$

where $z_i \in \mathbb{R}^N$ denotes the i^{th} row of $\mathbf{z}$. z_{ij} is the probability that data point i is representative for data point j, hence $z_{ij} \in [0, 1]$. The row sparsity regularized trace minimization algorithm is applied on matrix $\mathbf{Z}$ to select some representative exemplars that can suitably encode pose data according to dissimilarities as follows:

$$\min \sum_{j=1}^{N} \sum_{i=1}^{N} d_{ij} z_{ij} + \eta \sum_{i=1}^{N} \left\| z_i \right\|_q, \tag{4.3}$$

subject to:

$$z_{ij} \geq 0, \quad \forall i, j; \quad \sum_{i=1}^{N} z_{ij} = 1, \quad \forall j, \tag{4.4}$$

where the parameter $\eta > 0$ sets the trade-off between both terms.

Once this optimization problem (Eq. 4.3) has been solved, one can find the representative indices from the nonzero rows of $\mathbf{Z}$. The clustering of data points into K clusters, associated with K representatives, is obtained by assigning each data point to its closest representative. In particular, if $\{ i_1; \ldots ; i_q \}$ denote the indices of the representatives, data point j is assigned to the pose representative $\theta(j)$ such that $\theta(j) = \arg\min_{\ell \in \{i_1;\ldots;i_q\}} d_{\ell j}$. The pose angle of representative video ROIs of each cluster, referred as exemplar, is used as rendering parameter to generate synthetic face images with varying poses using an off-the-shelf 3D face model (Blanz & Vetter (2003); Tran *et al.* (2017a,b)). For each pose exemplar, $\mathbf{p}_j$, a set of lighting exemplars are then selected. Clusters that represent a greater number of generic samples should have a greater influence on classification. Here, a weight is assigned to each exemplar to indicate its importance, approximated based on its cluster size, $W_{ij} = n_{ij}/n$, where n_{ij} is the number of samples in each cluster and n is the number of generic samples. This selection strengthens those classes that are more representative in reconstructing a probe sample.

4.3.3 Face Synthesis

To generate representative synthetic ROIs, we employ the domain-specific 3DMM, a simplified version of the 3DMM in which the texture fitting of the original 3DMM is replaced with image mapping based on target domain information (Mokhayeri *et al.* (2019a)). With this technique, each still reference image, is decomposed and its material layer is extracted based on the a texture-aware image model defined by Jeon *et al.* (2014). 3D shape models of reference ROIs, $\mathbf{r}_i$, are reconstructed using the 3DMM and rendered w.r.t. pose exemplars. Basically, the shape model is defined as a convex combination of shape vectors of a set of examples in which the shape vector ($\mathbf{S}$) is defined as Eq.4.5 (Blanz & Vetter (2003)).

$$\mathbf{S} = \bar{\mathbf{S}} + \sum_{k=1}^{m_S-1} \alpha_k . \tilde{\mathbf{S}}_k \ , \tag{4.5}$$

where, the 3D shape is represented by the probability distribution of faces around the averages of shape $\bar{\mathbf{S}}$ are calculated and the basis vectors $\tilde{\mathbf{S}}_j$, $1 \leq j \leq m_s$ in Eq.4.5 where m_s is the number of the basis vectors. Here, for each reference ROI, $\mathbf{r}_i$, we reconstruct the 3D shape using:

$$\mathbf{S}_i = \bar{\mathbf{S}} + \sum_{j=1}^{m_S - 1} \alpha_j^i . \tilde{\mathbf{S}}_j, \tag{4.6}$$

where $\alpha_j^i \in [0,1], 1 \leq j \leq m_s$ are the shape parameters and $\mathbf{S}_i$ is the reconstructed shape of the i^{th} reference still ROI $\mathbf{r}_i$. The optimization algorithm presented by Blanz & Vetter (2003) is employed to find optimal $\alpha_j^i, 1 \leq j \leq m_s$, for each reference still ROI $\mathbf{r}_i$. The extracted material layers are then projected to the 3D model of the reference gallery set. Given the 3D facial shape and texture, novel poses can be rendered under various forms of the pose by adjusting the parameters of a camera model.

Following this, the shading layers of the lighting exemplars are projected on the rendered views by morphing between the layers. A guided filter with the guidance of the input shading layer is applied to preserve the structure of the input face. In this way, q synthetic faces, $\mathbf{S}_i = \{\mathbf{s}_1, \ldots, \mathbf{s}_q\} \in \mathbb{R}^{d \times q}$, are generated under the representative information from a given single still face image. The augmented gallery dictionary $\mathbf{D}' = \{\mathbf{R}'_1, \ldots, \mathbf{R}'_i, \ldots, \mathbf{R}'_k\} \in \mathbb{R}^{d \times k(n+q)}$ is formed by merging each still ROI of (k class) reference set with q synthetic images rendered w.r.t. representative exemplars, where $\mathbf{R}'_i = \{\mathbf{r}^i_1, \ldots, \mathbf{r}^i_n, \mathbf{s}_1, \ldots, \mathbf{s}_q\} \in \mathbb{R}^{d \times (n+q)}$.

4.3.4 Block-Sparsity Matching

After generating the representative synthetic samples, a Siamese network with ResNet-50 architecture (He *et al.* (2016)) is designed that learns how to differentiate between two inputs. A deep Siamese network basically consists two symmetrical CNN feature extractors both sharing the same weights and architecture, and both joined together at the end using some energy function. When a query, y is matched against a gallery image (either the reference face, r_i or synthetic faces, s_i of person i), the last layer of the CNNs produce fixed size vector, f_y, and f_r, and the SiamSRC network finally outputs a similarity scores. The proposed SiamSRC model employs sparse representation for pair-wise face matching (see Fig. 4.2). SRC derives the sparse coefficients $\boldsymbol{\alpha}$ of $\mathbf{y}$ by solving the following ℓ_1-minimization problem:

$$\min_{\alpha} \|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1. \tag{4.7}$$

where λ is a regularization parameter, and $\lambda > 0$. After the sparse vector of coefficients $\boldsymbol{\alpha}$ is obtained, the probe image $\mathbf{y}$ is recognized as belonging to class k^* if it satisfies:

$$k^* = \arg\min_{k} \|\mathbf{y} - \mathbf{D}\gamma_k(\boldsymbol{\alpha})\|_2. \tag{4.8}$$

where γ_k is a vector whose only nonzero entries that are the entries in $\boldsymbol{\alpha}$ are associated with class k. That is, the probe image $\mathbf{y}$ will be assigned to the class with the minimum class-wise reconstruction error.

Classification can rely on the reconstruction of a query ROI in the dictionary of all reference images, using the so-called structured SRC (S-SRC). Since the synthetic ROIs generated for each individual forms a block inside the gallery, higher classification accuracy is possible if the reconstruction of a query ROI is produced from the minimum number of blocks from the dictionary. The goal of S-SRC is to find a representation of a probe ROI that uses the minimum number of blocks from the dictionary. The block sparsity is formulated in terms of mixed ℓ_t/ℓ_1 ($t > 1$) norm as follows.

$$A_{\ell_2/\ell_1}: \quad \min_{\alpha} \|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{i=1}^{q} \|\boldsymbol{\alpha}\|_1. \tag{4.9}$$

where $\boldsymbol{\alpha}[i]$ is the i^{th} block in the sparse coefficient vector $\boldsymbol{\alpha}$ corresponding to the dictionary block $\mathbf{D}[i]$. Since each dictionary block corresponds to a specific class, i represents the class index ranging from 1 to q as well. This is a convex optimization problem when $t \geq 1$. Here we suppose $t = 2$. This optimization problem seeks the minimum number of non-zero coefficient blocks that reconstruct the probe ROI. Finally, the weighted matrix obtained in Section 4.3.2

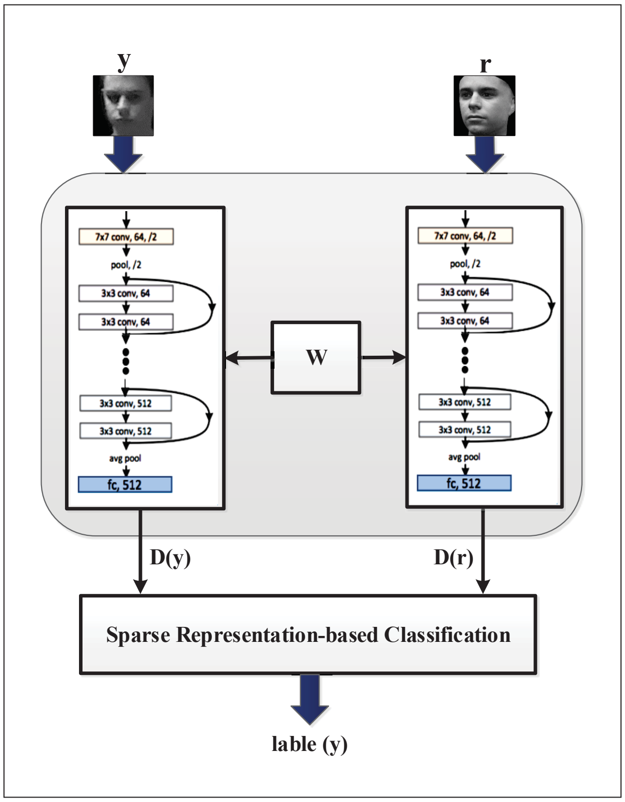


Figure 4.2 Architecture of SiamSRC.

shows that cluster weights multiplied to the ℓ_1-minimization term.

$$A_{\ell_2/\ell_1}: \quad \min_{\alpha} \|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{i=1}^{q} \|W\boldsymbol{\alpha}\|_1. \tag{4.10}$$

where

$$\mathbf{W}_i = \begin{bmatrix} W_{i1} & 0 & \cdots & 0 \\ 0 & W_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_{iq} \end{bmatrix} \tag{4.11}$$

The class label of the probe ROIs y is then determined based on the block sparse reconstruction error as 4.12. Algorithm 4.1 summarizes the steps for design and testing with the SiamSRC network.

$$\text{label}(\mathbf{y}) = \arg\min_{k} \| \mathbf{y} - \mathbf{D}[k]\gamma_k \boldsymbol{\alpha}[k] \|_2 \ . \tag{4.12}$$

Algorithm 4.1 SiamSRC Network.

```
Input: Reference still ROIs $\{\mathbf{R}_i : i = 1,\dots,n\}$, Generic set of video ROIs $\{\mathbf{G}_i : i = 1,\dots,M\}$, a
       probe sample $\mathbf{y}$, and parameters $\lambda$, $\mu$, and $\xi$
Design:
Estimate pose angles and luminance-contrast distortions of ROIs in $\{\mathbf{G}_i : i = 1,\dots,M\}$
Cluster video ROIs in the pose and lighting space with the row sparsity technique, and produce $q$
  representative exemplars
for each still $\mathbf{R}_i$ do
    Generate $q$ synthetic images $\{\mathbf{S}_i : i = 1,\dots,q\}$ per each individual based on exemplars
    Merge $\{\mathbf{S}_i : i = 1,\dots,q\}$ with $\{\mathbf{R}_i : i = 1,\dots,n\}$ to form $\{\mathbf{R'}_i : i = 1,\dots,(n+q)\}$
    Integrate augmented gallery $\mathbf{R'}_i$ into the SiamSRC network
end
Testing:
Match faces by solving the sparse representation problem to estimate coefficient matrix, $\boldsymbol{\alpha}$ for $\mathbf{y}$
  (see Eq. 4.10)
Output: $label(\mathbf{y}) = \arg\min_{\mathbf{k}}(\mathbf{r_k}(\mathbf{y}))$ (see Eq. 4.12)
```

The main steps of the proposed domain-invariant S2V FR with dictionary augmentation are summarized as follows.

Selecting representatives: During design, a generic set $\mathbf{G}_i \in R^{d \times M}$ in the target domain is clustered using row sparsity in the pose angles space.

Augmentation of the reference gallery: q synthetic ROIs, $\mathbf{S}_i \in R^{d \times q}$ are generated for each $\mathbf{r}_i$ of the reference gallery set in the source domain to form an augmented reference gallery $\mathbf{R'}_i \in \mathbb{R}^{d \times k(q+1)}$, where q is the number of clusters.

Block-sparsity matching: During inference, a given an input query image, $\mathbf{y}$, in matched to a set of images in the augmented gallery, $\mathbf{R'}_i \in \mathbb{R}^{d \times k(q+1)}$. For each matching, the Siamese network outputs a deep feature representation, SRC with block structure is used to compute similarity scores between the features. Figure 4.3 summarizes a typical implementation of the SiamSRC network for still-to-video FR.

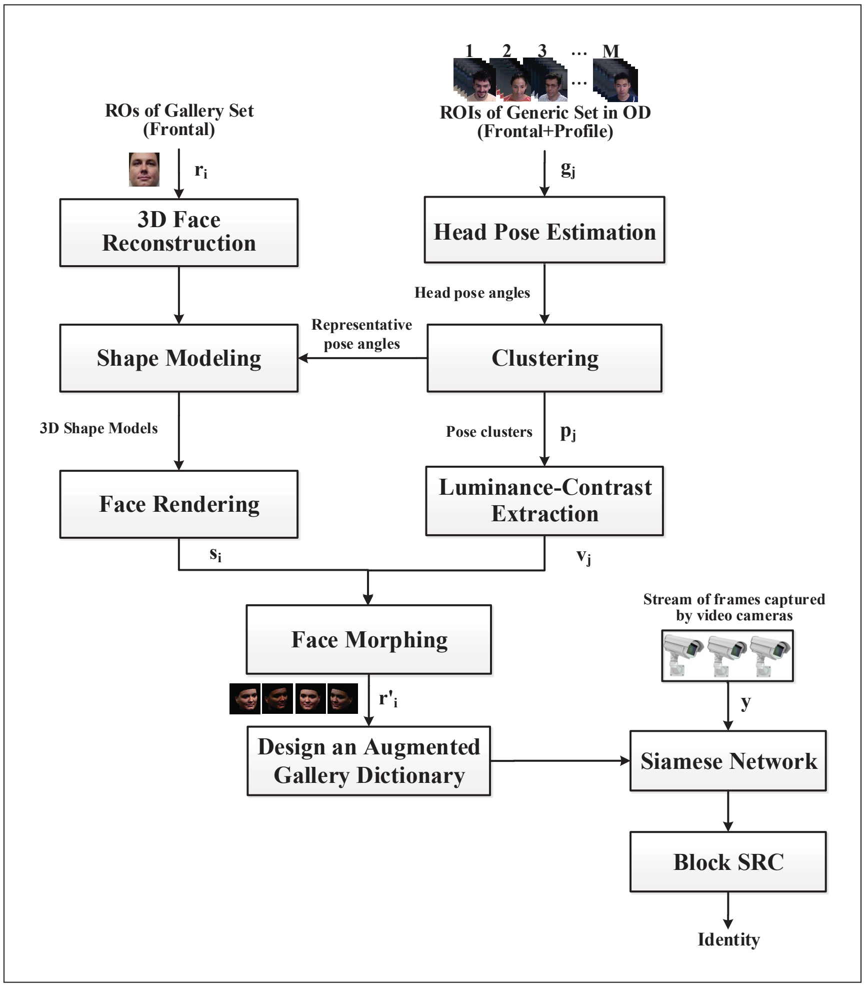


Figure 4.3 The block diagram that summarizes the steps for design and inference with the proposed SiamSRC network.

4.4 Experimental Methodology

4.4.1 Datasets

In order to evaluate the accuracy and complexity of the proposed SiamSRC network, an extensive series of experiments are conducted on publicly-available datasets for still-to-video FR. Datasets to validate still-to-video FR systems should contain at least one good quality reference frontal still per person, and many lower-quality video sequences captured under various uncontrolled conditions, incorporating real-world intra-class variations. The Chokepoint[1] (Wong *et al.* (2011)) and COX-S2V[2] (Huang *et al.* (2015)) datasets are suitable for experiments in still-to-video FR in video surveillance because they are composed of a high-quality still image and several lower-resolution video sequences per subject, with variations of illumination, pose, expression, motion blur, and scale, etc. Chokepoint is comprised of a smaller number subjects, while the COX-S2V dataset is comprised of 1,000 subjects.

Chokepoint (Wong *et al.* (2011)) consists of 25 subjects walking through portal 1 and 29 subjects in portal 2. Videos are recorded over 4 sessions one month apart. An array of 3 cameras are mounted above P1 and P2 that capture the subjects during 4 sessions while they are either entering or leaving the portals in a natural manner. In total, 4 data subsets are available, and the dataset consists of 54 video sequences.

COX-S2V dataset (Huang *et al.* (2015)) contains $1,000$ individuals, with 1 high-quality still image and $3,000$ low-resolution video sequences per each individual simulating video surveillance scenario. The video frames are captured by 4 cameras mounted at fixed locations of about 2 meters high. In each video, an individual walk through an S-shape route with changes in illumination, scale, and pose.

4.4.2 Protocol and Performance Measures

A particular implementation of the SiamSRC model has been considered to validate the proposed approach. We add synthetic facial images to the gallery of the Siamese network, and employ SRC for the pair-wise face matching. First, during the calibration process, q representatives are selected based on the q clusters obtained using the row sparsity clustering on facial ROI trajectories of unknown persons captured in the target domain. During the enrollment of an individual to the system, the ROIs of the generic set of faces captured from the video trajectories are extracted using the MTCNN face detection algorithm (Zhang *et al.* (2016)) and

[1] http://arma.sourceforge.net/chokepoint.

[2] http://vipl.ict.ac.cn.

converted to grey-scale images of 96×96 pixels. Pre-processing is also applied to still images prior to face synthesis. q synthetic ROIs for each reference still are generated. The augmented gallery is constructed using the reference still ROIs along with their synthetic ROIs. The structure of the network is ResNet-50 which is pre-trained with the images of the Labeled Faces in the Wild (LFW) dataset [3] (Huang *et al.* (2012)) which is a suitable dataset for large scale face recognition in the wild. The dataset contains more than 13,000 images of faces collected from the web. 1680 of the people pictured have two or more distinct photos in the dataset, and each face has been labeled with the name of the person pictured. The sparsity parameter λ is fixed to 0.005 during the experiments. Each probe or query image is presented to the network to obtain a deep feature representation, and then block sparse coding is used for pair-wise face matching. In all experiments with Chokepoint dataset, 5 target individuals are selected randomly to design a watch-list that includes a high-quality frontal captured images, and for the experiment with COX-S2V, 20 individuals are randomly selected to build a watch-list from high-quality faces. Videos of 10 individuals that are assumed to come from non-target persons are used as generic set. The rest of the videos including 10 other non-target individuals and the videos of individuals who are already enrolled in the watch-list are used for testing. In order to obtain representative results, this process is repeated 5 times with a different random selection of watch-lists and the average performance is reported with standard deviation over all the runs.

The average accuracy of the proposed and baseline FR systems are measured in terms of accuracy and complexity. For accuracy, we measure the partial area under ROC curve pAUC(20%) (using the AUC at $0 < FPR \leq 20\%$), and the area under precision-recall space (AUPR). Time complexity is estimated empirically, using the amount of time required to match 2 facial ROIs with the given dataset. The average running time is measured with a randomly selected probe ROIs using a PC workstation with an Intel Core i7 CPU (3.41GHz) processor, 16GB RAM and python, tensorflow (GPU version).

[3] http://vis-www.cs.umass.edu/lfw/.

4.5 Results and Discussion

This section first shows representative selection results and some examples of synthetic face images, and then presents still-to-video FR performance achieved with augmenting the gallery with such images. In order to investigate the impact of the proposed SiamSRC network on performance, we considered the still-to-video FR system with a growing number of synthetic faces, along with a generic training set. Finally, this section presents an ablation study (showing the effect of each module on the performance) and a complexity analysis for our proposed approach and several state-of-the-art methods: S+V model (Mokhayeri & Granger (2018)), DeepFace (Taigman *et al.* (2014)), FaceNet (Schroff *et al.* (2015)), ESRC (Deng *et al.* (2012)), SVDL (Yang *et al.* (2013)), RADL (Wei & Wang (2015)), LGR (Zhu *et al.* (2014)), CSR (Li *et al.* (2016)), and face frontalization (Hassner *et al.* (2015))..

4.5.1 Synthetic Face Generation

Fig. 4.4 shows an example of the row sparsity clustering obtained with facial ROIs of 20 trajectories extracted from Chokepoint videos of 5 individuals, and of 40 trajectories extracted from COX-S2V videos of 10 individuals. The exemplars selected from these clusters (black circles) are used to define rendering parameters for synthetic generation. For instance, the representative pose angles with the Chokepoint database are listed as follows: $\theta_{Chok1} =$ (pitch, yaw, roll)= (15.65, 14.77, -0.62), $\theta_{Chok2} =$ (12.44, 2.76, 3.64), $\theta_{Chok3} =$ (9.06, -5.46, 4.73), $\theta_{Chok4} =$ (1.98, 6.09, 2.79), $\theta_{Chok5} =$ (13.21, 15.32, 6.14), $\theta_{Chok6} =$ (0.64, -18.93, 0.86), $\theta_{Chok7} =$ (5.23, 2.92, 2.03) degrees. Another clustering is then applied in the illumination measure space on each pose cluster. Fig. 4.5 shows the synthetic face images generated based on domain-specific 3DMM under representative pose angles using reference still ROIs of the Chokepoint and COX-S2V datasets.

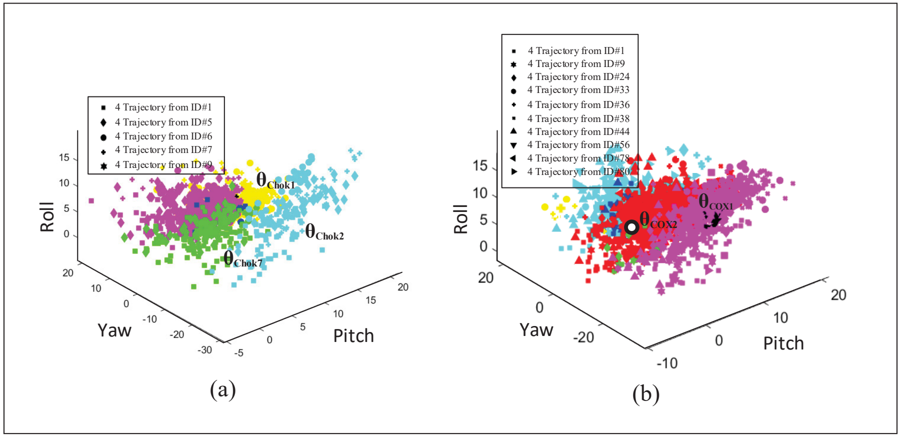


Figure 4.4 Example of clusters obtained with 20 and 40 facial trajectories represented in the pose space with Chokepoint (a) and COX-S2V (b) datasets, respectively. Clusters are shown with different colors, and their representative exemplars are indicated with a black circle.

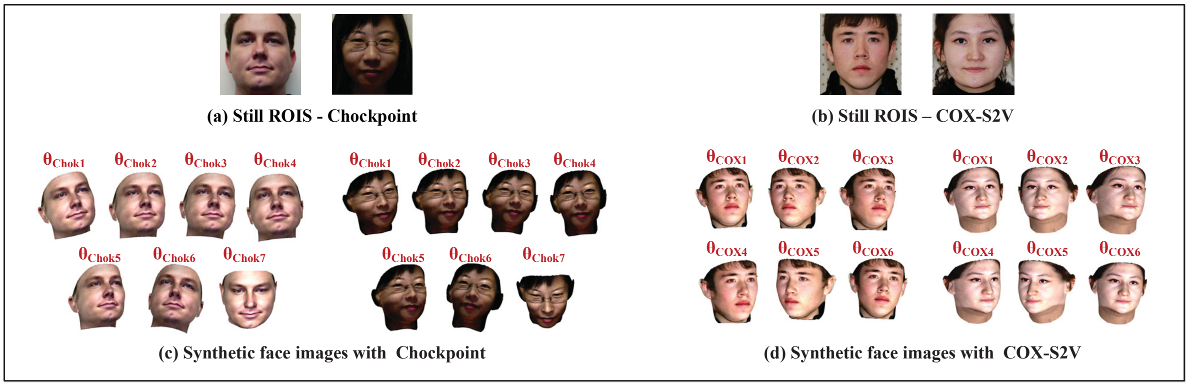
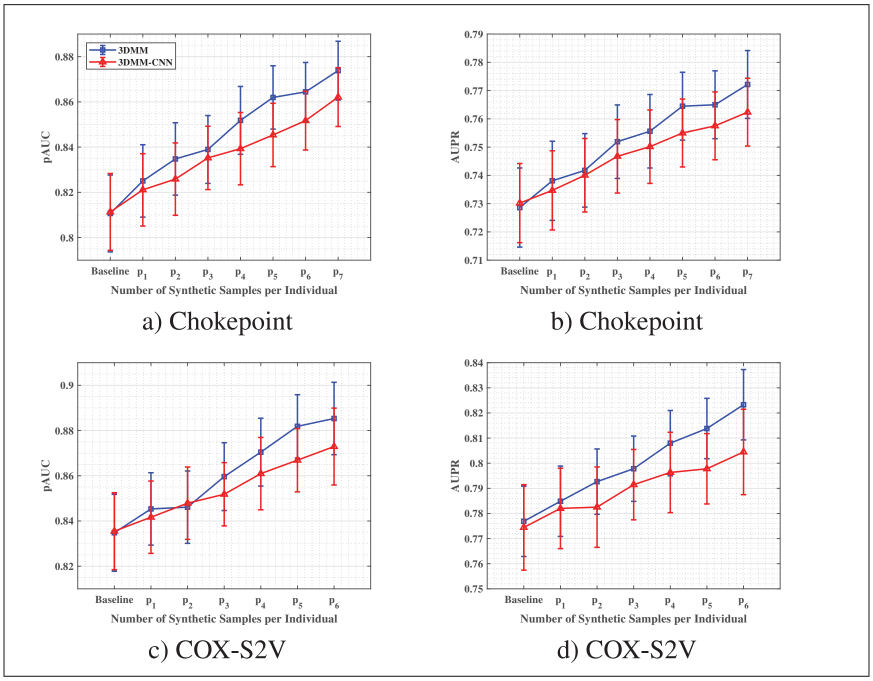

(a) Still ROIS - Chockpoint
(b) Still ROIS – COX-S2V
θ_{Chok1} θ_{Chok2} θ_{Chok3} θ_{Chok4}
θ_{Chok5} θ_{Chok6} θ_{Chok7}
θ_{Chok1} θ_{Chok2} θ_{Chok3} θ_{Chok4}
θ_{Chok5} θ_{Chok6} θ_{Chok7}
θ_{COX1} θ_{COX2} θ_{COX3}
θ_{COX4} θ_{COX5} θ_{COX6}
θ_{COX1} θ_{COX2} θ_{COX3}
θ_{COX4} θ_{COX5} θ_{COX6}
(c) Synthetic face images with Chockpoint
(d) Synthetic face images with COX-S2V

Figure 4.5 Examples of the reference still ROI of individuals ID#25 and ID#26 of Chokepoint dataset (a), ID#21 and ID#151 of COX-S2V dataset (c), and their corresponding synthetic face images (b,d) produced using the representatives and domain-specific 3DMM.

4.5.2 Impact of Number of Synthetic Images

Fig. 4.6 shows the average pAUC(20%) and AUPR accuracy obtained by the SiamSRC network when increasing the number of synthetic ROIs under representative pose per each individual. These ROIs were sampled from the q representatives exemplars selected from the Chokepoint and COX-S2V datasets. We also show results for 3DMM-CNN technique (Tran *et al.* (2017a)) that uses a CNN to regress 3DMM shape parameters directly from an input photo without an optimization process. Fig. 4.7 shows the average pAUC(20%) and AUPR accuracy of the SiamSRC network when increasing the number of synthetic ROIs under various pose and illumination conditions in the augmented gallery.



a) Chokepoint

b) Chokepoint

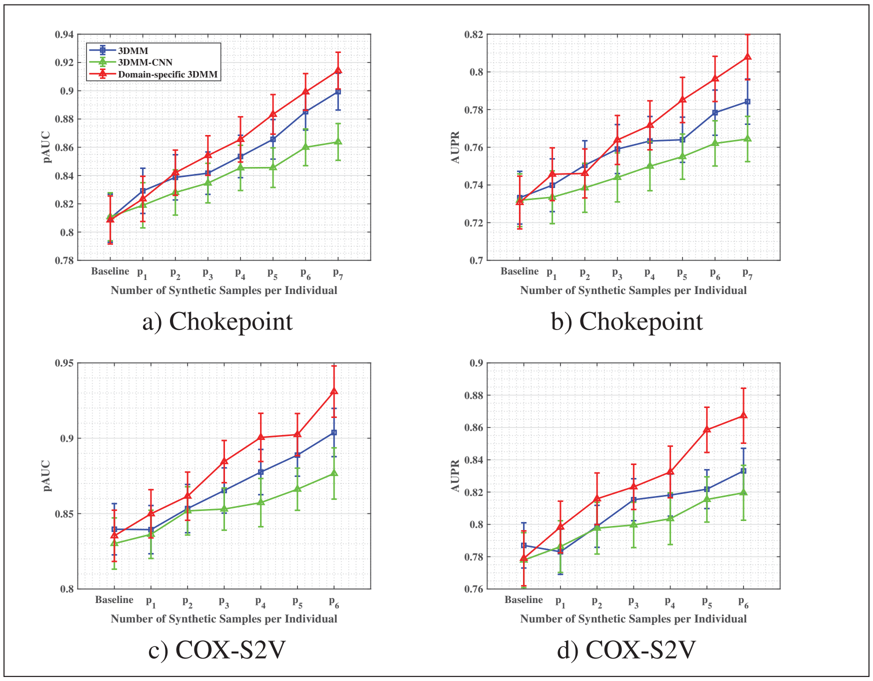
c) COX-S2V

d) COX-S2V

Figure 4.6 Average pAUC(20%) and AUPR accuracy of SiamSRC network versus the size of the synthetic set on Chokepoint (a,b) and COX-S2V (c,d) databases. Synthetic faces are generated using 3DMM, domain-specific 3DMM, and 3DMM-CNN. Error bars are standard deviation.

The results indicate that adding representative synthetic ROIs to the reference gallery allows to outperform the baseline system designed with an original reference still ROI alone. AUC and AUPR accuracy increase considerably by about $20-30\%$ with only $q_{Chok}=7$ and $q_{COX}=6$ synthetic pose ROIs for Chokepoint and COX-S2V datasets, respectively.

To further assess the benefits, Fig. 4.8 compares the performance of the proposed SiamSRC network when adding representative images versus 15 randomly selected images. Results in



a) Chokepoint

b) Chokepoint

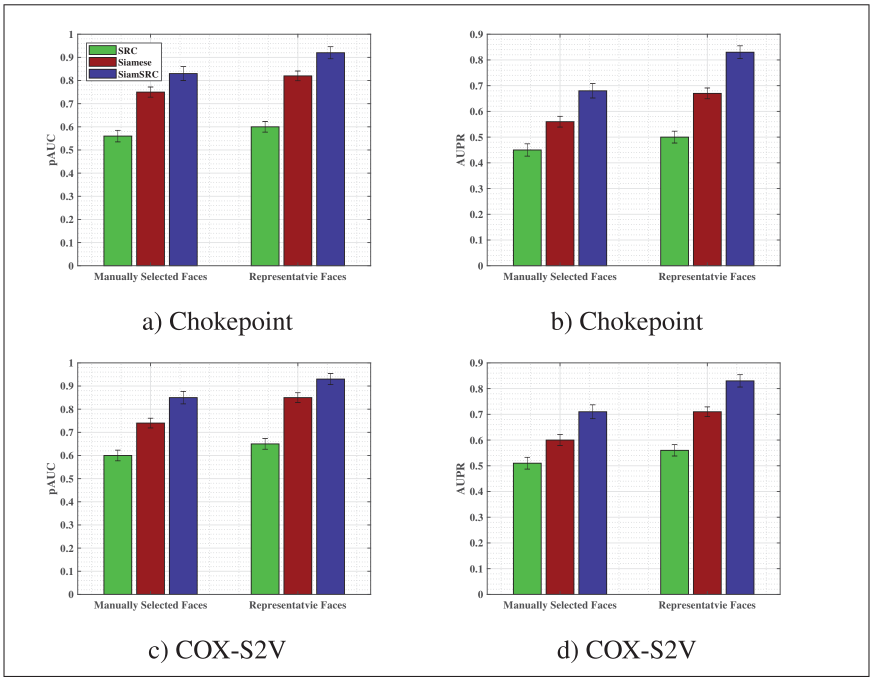
c) COX-S2V

d) COX-S2V

Figure 4.7 Average pAUC(20%) and AUPR accuracy of SiamSRC network versus the size of the synthetic set on Chokepoint (a,b) and COX-S2V (c,d) databases. Synthetic faces are generated using 3DMM, domain-specific 3DMM, and 3DMM-CNN. Error bars are standard deviation.

this figure show that SiamSRC outperforms the two other models – FR performance is higher when the gallery is designed using the q representative views than based gallery comprised of 15 randomly selected synthetic faces per person. The proposed SiamSRC network can therefore adequately generate representative facial ROIs for the gallery.

The results of SiamSRC and some generic learning techniques can be evaluated based on the size of the generic set. Given N generic images (video ROIs) from the target domain, the recognition rate of the approaches is compared with increasing of N. Fig. 4.9 shows that for several generic learning techniques, the intra-class variation information of a small number of individuals is sufficient to largely improve the recognition rate. In particular, it can be observed from Fig.4.9 that when more generic images are used, the accuracy increases significantly for



a) Chokepoint

b) Chokepoint

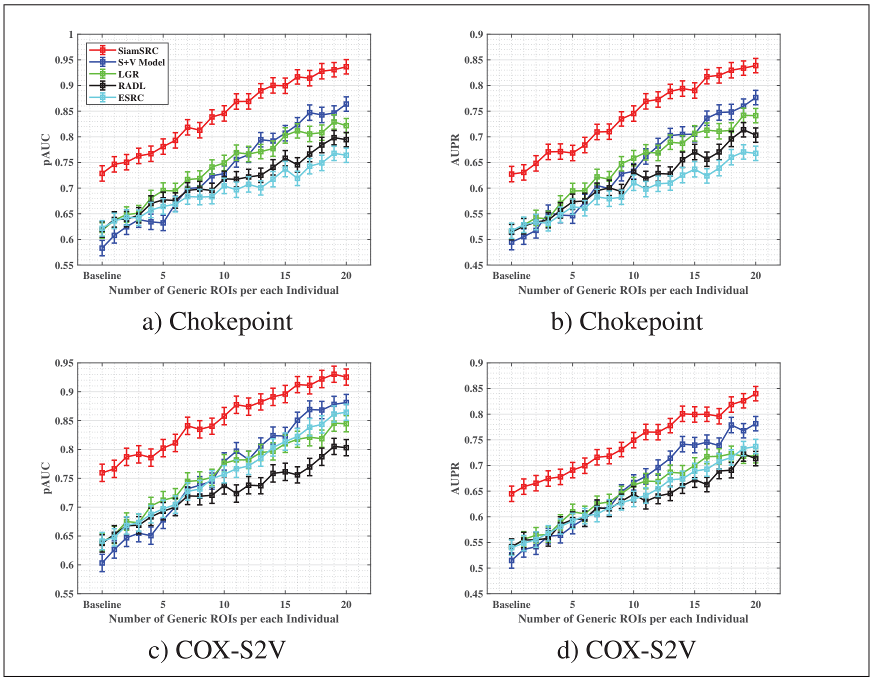
c) COX-S2V

d) COX-S2V

Figure 4.8 Average pAUC(20%) and AUPR accuracy for SRC model, and Siamese and SiamSRC networks with *q* representative and 15 randomly selected faces on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation.

SiamSRC and S+V techniques. This shows that the proposed SiamSRC method is able to properly select representative faces out of a set of faces.

Fig. 4.10 shows the impact on SiamSRC accuracy of adding generic set and synthetic faces. The performance of the S+V model is also shown with the same configurations for the comparison. The results indicate that the accuracy of the SiamSRC network improves by adding synthetic samples, but does not change with adding generic set faces. However, the performance of S+V model increases significantly with the inclusion of both synthetic faces and generic set. The reason why SRC techniques can benefit from generic set is that they are able to decompose the signal as a linear combination of few signals, as a result, they can recover a probe signal as a combination of variation of generic set and original gallery signal.



a) Chokepoint

b) Chokepoint

c) COX-S2V

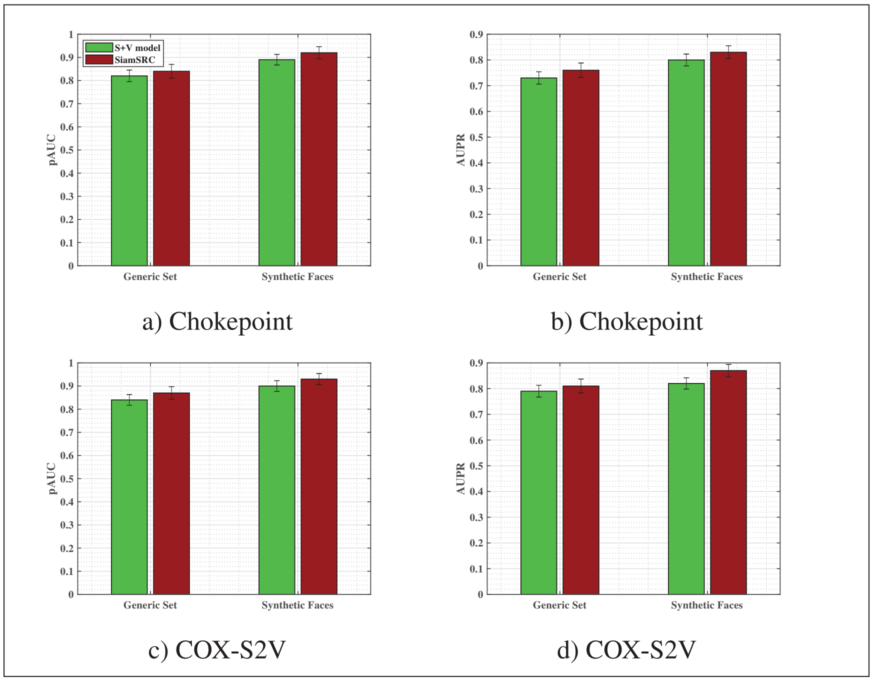
d) COX-S2V

Figure 4.9 Average pAUC(20%) and AUPR accuracy versus the size of the generic set on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation.

4.5.3 Comparison with State-of-the-Art Methods

Table 4.1 compares the FR accuracy obtained with the proposed SiamSRC network with baseline methods: Original Siamese (Deng *et al.* (2012)), Deep Face (Deng *et al.* (2012)), FaceNet (Deng *et al.* (2012)), and SRC techniques: S+V model (Mokhayeri & Granger (2018)), ESRC (Deng *et al.* (2012)), SVDL (Yang *et al.* (2013)), LGR (Zhu *et al.* (2014)), RADL (Wei & Wang (2015)), CSR (Li *et al.* (2016)). We also compared it with the Flow-Based Face Frontalization (Hassner *et al.* (2015)) technique. The results shows that the SiamSRC model, using a joint generic learning and face synthesis, achieves the higher level of accuracy under the same configuration.

Table 4.2 shows the average matching time of deep Siamese networks over the video ROIs of the Chokepoint and COX-S2V datasets. The table shows time complexity increases with the growth of the gallery size. Since SiamSRC requires a moderate increase in computational



a) Chokepoint

b) Chokepoint

c) COX-S2V

d) COX-S2V

Figure 4.10 Average pAUC(20%) and AUPR accuracy for S+V model and SiamSRC network of using an augmented gallery with both generic set and synthetic faces on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation.

complexity w.r.t. to other Siamese networks, results suggest that the proposed approach is able to represent an interesting trade-off between accuracy and complexity. In the future, we plan to validate the proposed approach with the IJB-A dataset (Whitelam *et al.* (2017)) which consists of a larger number of stills, videos, and subjects.

4.5.4 Ablation Study

Designing the SiamSRC network for still-to-video FR consists of three main steps: selection of representatives ($\mathcal{M}_1$), face synthesis ($\mathcal{M}_2$), and SRC for face matching ($\mathcal{M}_3$). In this subsection, an ablation study is presented to show the impact of each module on FR performance. We assume that all FR systems use domain-specific 3DMM face synthesis, and q synthetic images in the augmented dictionary. First, we disabled the representative selection module, $\mathcal{M}_1$, and added 10 randomly selected synthetic ROIs. Next, we removed the face synthesis mod-

Table 4.1 Average accuracy of the SiamSRC network and related state-of-the networks for videos from all 3 cameras of the Chokepoint and COX-S2V databases. Feature representations are raw pixels, the 3DMM method is used for face synthesis.

Techniques	**Accuracy**			
	Chokepoint database		**COX-S2V database**	
	pAUC(20%)	AUPR	pAUC(20%)	AUPR
SRC (Wright *et al.* (2009))	0.524±0.032	0.475±0.031	0.568±0.030	0.480±0.027
ESRC (Deng *et al.* (2012))	0.802±0.028	0.651±0.025	0.835±0.027	0.695±0.026
ESRC-KSVD	0.811±0.023	0.659±0.022	0.840±0.023	0.712±0.021
Face Frontalization (Hassner *et al.* (2015))	0.822±0.021	0.711±0.023	0.843±0.022	0.719±0.023
SVDL (Yang *et al.* (2013))	0.825±0.023	0.703±0.025	0.843±0.025	0.724±0.023
RADL (Wei & Wang (2015))	0.832±0.019	0.711±0.020	0.849±0.022	0.730±0.021
LGR (Zhu *et al.* (2014))	0.849±0.022	0.717±0.024	0.878±0.023	0.744±0.025
CSR (Li *et al.* (2016))	0.852±0.025	0.722±0.020	0.880±0.021	0.753±0.020
S+V model (Mokhayeri & Granger (2018))	0.882±0.018	0.745±0.019	0.895±0.020	0.766±0.017
DeepFace (Taigman *et al.* (2014))	0.895±0.021	0.802±0.023	0.906±0.022	0.812±0.024
Cosface (Wang *et al.* (2018a))	0.903±0.022	0.810±0.020	0.907±0.025	0.816±0.023
Siamese Network (Koch *et al.* (2015)Schroff *et al.* (2015))				
· 1 frontal reference still / person	0.833±0.028	0.742±0.031	0.852±0.029	0.791±0.027
· block sparsity match	0.851±0.029	0.752±0.027	0.866±0.026	0.798±0.028
· 1 still+73 uniform synthetic / person	0.872±0.023	0.772±0.021	0.893±0.020	0.813±0.022
· 1 still+100 random synthetic / person	0.861±0.022	0.764±0.020	0.878±0.021	0.802±0.023
SiamSRC (Ours)				
· 1 still+q representative synthetic / person	**0.911±0.019**	**0.819±0.020**	**0.923±0.018**	**0.837±0.016**

Table 4.2 Average matching time of deep Siamese networks over the video ROIs of the Chokepoint and COX-S2V datasets.

Techniques	**Matching Time (sec)**	
	Chokepoint	**COX-S2V**
Siamese Network		
· 1 frontal reference still / person	0.07	0.09
· 1 still + 73 uniform synthetic / person	0.13	0.18
· 1 still + 100 random synthetic / person	0.15	0.21
SiamSRC (Ours)		
· 1 still + q representative synthetic / person	0.27	0.33

ule and performed experiments to show the impact of the reference gallery augmenting with synthetic faces on FR accuracy. By removing both $\mathcal{M}_1$ and $\mathcal{M}_2$ modules from the SiamSRC network, accuracy declines significantly by about 50%. The results suggest that the addition of representative synthetic faces is an effective strategy to cope with facial variations and prevent overfitting. Another important component of the SiamSRC model is the face matching through SRC. By removing the sparsity in SiamSRC model, $\mathcal{M}_3$, and replacing Euclidean distance, accuracy decreases by about 15%. Tables 4.3-4.4 show the average AUC and AUPR accuracy of the ablation study using the videos from all cameras of the Chokepoint and COX-S2V datasets, respectively. respectively.

Table 4.3 Ablation study results with Chokepoint dataset.

Accuracy	**Removed Module**			
	baseline	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
pAUC	0.911±0.019	0.852±0.21	0.846±0.27	0.893±0.25
AUPR	0.819±0.020	0.721±0.23	0.725±0.25	0.752±0.22

Table 4.4 Ablation study results with COX-S2V dataset.

Accuracy	**Removed Module**			
	baseline	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
pAUC	0.923±0.018	0.871±0.22	0.857±0.24	0.911±0.20
AUPR	0.837±0.016	0.745±0.20	0.739±0.21	0.782±0.21

4.6 Conclusion

This paper proposes the SiamSRC technique to improve the performance of still-to-video face recognition systems when surveillance videos are captured under various uncontrolled conditions, and individuals must be recognized based on a single facial still. The proposed approach exploits a deep Siamese network and sparse representation-based classification with block structure for pair-wide face matching. It also leverages domain-specific face synthesis where rendering parameters are obtained through row sparsity clustering of unlabeled faces captured in the target domain. Experimental results obtained using the Chokepoint and COX-S2V datasets suggest that the proposed SiamSRC network allows for efficient representation of intra-class variations with only a moderate increase in time complexity. Results indicated that performance of still-to-video FR systems based on SiamSRC can improve through face synthesis, with no need to collect a large amount of training data. Future research includes investigating the geometrical structure of the data space in the gallery and the corresponding coefficients to improve discrimination. Future work plan to speed-up the proposed SiamSRC network and represent larger sparse codes which is necessary for real-world surveillance applications.

CONCLUSION AND RECOMMENDATIONS

Still-to-video face recognition has become more significant in recent years owing to its potential applications as found in surveillance and security. It offers remarkable advantages over other biometric modalities such as convenient data collection, easy installation, and flexible control. Although progress in face recognition has been encouraging in the past few years, many problems still exist in unconstraint tasks, especially when the size of training reference still images is small. This thesis aimed to design a robust still-to-video face recognition system by improving representative ability through enlarging the reference gallery synthetically. In this way, we proposed face synthesizing methods that generate images under target domain capture conditions by exploiting the discriminant information of the generic set.

In Chapter 2, we presented a domain-specific face synthesis algorithm by integrating an image-based face relighting technique inside a 3D face reconstruction framework. With the proposed technique, a compact set of synthetic faces are generated that represent reference images and probe video frames under a common capture condition. Results indicate that face synthesis alone cannot effectively resolve the challenges of the SSPP and visual domain shift problems for still-to-video FR. Integrating generic learning with face synthesis in the domain specific context lead to desire performance.

In Chapter 3, we proposed a paired sparse representation framework to account for linear and non-linear variations in the context of still-to-video FR. The proposed model leveraged both face synthesis and generic learning to effectively represent probe ROIs from a single reference still with no need to collect a large amount of training data. This approach managed the non-linear variations by enriching the gallery dictionary with a representative set of synthetic profile faces, where synthetic faces are paired with generic set video face in the auxiliary variational dictionary. Results show that generic learning alone cannot effectively resolve the challenges of the SSPP and visual domain shift problems.

In Chapter 4, we presented the SiamSRC network that employs block sparsity for face matching inside a Siamese network. In order to improve its robustness, a compact set of domain-specific facial images is further generated and integrated into the augmented reference gallery of the Siamese network. Results suggest that the proposed SiamSRC network allows for efficient representation of intra-class variations and can improve the performance of still-to-video FR systems accordingly.

Lastly, we introduced the controllable GAN that generates identity preserving and realistic synthetic faces under specific pose. The proposed technique extends the original GAN by using synthetic images as inputs instead of random noise vector. It also employs an additional adversarial game as a third player to provide control over the face generation process. Result indicate that the synthetic face images generated based on the controllable GAN allow us address visual domain shift, and thereby improve the accuracy of still-to-video FR system with no need to generate a large number of synthetic face images.

It can be concluded that face synthesis methods proposed in this thesis not only are able to generate realistic synthetic face images under target domain capture conditions but also control the conditions under which synthetic faces are generated. They can appropriately address visual domain shift and single sample per person problems for still-to-video FR applications with only a moderate increase in their computational complexity. They provide a higher level of accuracy compared to the current state-of-the-art approaches for synthetic data augmentation.

Future Works

The findings of this thesis suggest the following directions for future work in this topic:

- **Learning the gap between current 3DMM renderings and real-world 2D images.** Images generated by 3DMM usually lack facial details such as wrinkles or moles which are challenging to render properly. Although generative adversarial networks aims to address those challenges as texture models but they are not modeled in the shape and the resulting models lack details.

- **Generating synthetic images under a wide variety of facial appearance.** Although our proposed face synthesis algorithms have effectively improved recognition performance under unconstrained capture conditions, fundamental challenges such as matching faces cross ages, expressions, sensors, or styles still remain which are necessary to address in future research.

- **Design the proposed face synthesis model for the situations where we are not able to collect paired data in target domain.** Since it is often difficult to collect a large amount of generic set, the ability to use unpaired data with high accuracy enables us to avoid sophisticated and expensive paired data collection.

- **Reducing the reconstruction time**. We aim to reduce the reconstruction time of the proposed paired sparse and SiamSRC models even further allowing them to represent larger sparse codes. Because time complexity is the critical part of designing face recognition systems for real-world surveillance applications.

LIST OF PUBLICATIONS

Journal Articles

- Mokhayeri, F., Granger, E., and Bilodeau, G-A. "Domain-specific face synthesis for video face recognition from a single sample per person." IEEE Transactions on Information Forensics and Security, 14.3 (2018): 757-772.
- Mokhayeri, F., Granger, E. "Video face recognition using siamese networks with block-sparsity matching." IEEE Transactions on Biometrics, Behavior, and Identity Science (2019).
- Mokhayeri, F., Granger, E. "A paired sparse representation model for robust face recognition from a single sample." Pattern Recognition, 100 (2020): 107-129.

Conference Papers

- Mokhayeri, F., Eric Granger, and Bilodeau, G-A. "Synthetic face generation under various operational conditions in video surveillance." In 2015 IEEE International Conference on Image Processing (ICIP), 2015.
- Mokhayeri, F., Granger, E. "Robust video face recognition from a single still using a synthetic plus variational model." In 2019 IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2019.
- Mokhayeri, F., Granger, E. "Cross-Domain Face Synthesis using a Controllable GAN." In 2020 IEEE International Winter Conference on Applications of Computer Vision (WACV), 2020.

APPENDIX I

CROSS-DOMAIN FACE SYNTHESIS USING A CONTROLLABLE GENERATIVE ADVERSARIAL NETWORK

Fania Mokhayeri, Eric Granger

Le Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA),
École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article published in « Proceedings of the International Winter Conference on Applications of Computer Vision (WACV)» 2020.

Abstract

The performance of face recognition (FR) systems applied in video surveillance has been shown to improve when the design data is augmented through synthetic face generation. This is true, for instance, with pair-wise matchers (e.g., deep Siamese networks) that typically rely on a reference gallery with one still image per individual. However, generating synthetic images in the source domain may not improve the performance during operations due to the domain shift w.r.t. the target domain. Moreover, despite the emergence of Generative Adversarial Networks (GANs) for realistic synthetic generation, it is often difficult to control the conditions under which synthetic faces are generated. In this paper, a cross-domain face synthesis approach is proposed that integrates a new Controllable GAN (C-GAN). It employs an off-the-shelf 3D face model as a simulator to generate face images under various poses. The simulated images and noise are input to the C-GAN for realism refinement which employs an additional adversarial game as a third player to preserve the identity and specific facial attributes of the refined images. This allows generating realistic synthetic face images that reflects capture conditions in the target domain while controlling the GAN output to generate faces under desired pose conditions. Experiments were performed using videos from the Chokepoint and COX-S2V datasets, and a deep Siamese network for FR with a single reference still per person. Results

indicate that the proposed approach can provide a higher level of accuracy compared to the current state-of-the-art approaches for synthetic data augmentation[1].

2. Introduction

Recent advances in deep learning have significantly increased the performance of still-to-video face recognition (FR) systems applied in video monitoring and surveillance. One of the pioneering techniques in this area is FaceNet (Schroff *et al.* (2015)). It uses a deep Siamese network architecture, where the same CNN feature extractor is trained through similarity learning to perform pair-wise matching of query (video) and reference (still) faces. Despite many recent advances, FR with a single sample per person (SSPP) remains a challenging problem in video-based security and surveillance applications. In such cases, the performance of deep learning models for FR can decline significantly due to the limited robustness of matching to a single still obtained during enrolment (Bashbaghi *et al.* (2017a)). One effective solution to alleviate the aforementioned problem is extending the gallery using synthetic face images.

Some of the recent research (Mokhayeri *et al.* (2019a); Zhao & et al. (2017)) augment galleries using synthetic images generated from 3D models. Tran *et al.* (2017a) proposed a face synthesis technique where CNN is employed to regress the 3D model parameters to overcome the shortage of training data. Although their results are encouraging, the synthetic face images may not be realistic enough to represent intra-class variations of target domain capture conditions. The synthetic images generated in this way are highly correlated with the original facial stills from enrolment, and there is typically a domain shift between the distribution of synthetic faces and that of faces captured in the target domain. The models naively trained on these synthetic images, often fail to generalize well when matched to real images captures in the target domain. Mokhayeri *et al.* (2019a) proposed an algorithm for domain-specific face synthesis (DSFS) that exploits the intra-class variation information available from the target domain.

[1] Code available at: https://github.com/faniamokhayeri/C-GAN.

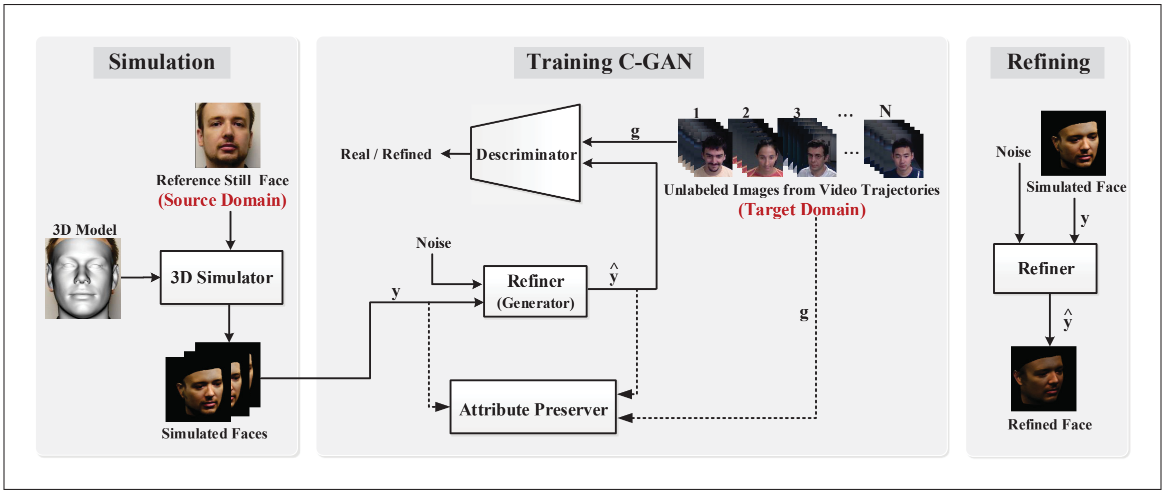


Figure-A I-1 An overview of the proposed cross-domain face synthesis approach based on the C-GAN. The 3D simulator generates simulated faces, **y**, with the arbitrary pose. The refiner is trained using **y** the generic set, **g**, and random noise to generate refined images, $\hat{\mathbf{y}}$, under the target domain capture conditions, and while specifying the pose of **y** using an additional adversarial game.

Generative adversarial networks (GANs) have recently shown promising results for the synthesis of realistic face images (Brock *et al.* (2019); Goodfellow & et al. (2014)). For instance, DA-GAN has been proposed for automatically generating augmented data for FR in unconstrained conditions (Zhao & et al. (2017)). One of the challenging issues in GAN-based face synthesis models is the difficulty controlling images they generate since a random distribution is used as the input of generators. Modified GAN architectures, like the conditional GAN, have attempted to address this issue by setting conditions on the generative and discriminative networks for conditional image synthesis (Isola *et al.* (2017); Lin *et al.* (2018b); Tran *et al.* (2018)). However, the mapping of conditional GANs does not constrain the output to the target manifold, thus the output can be arbitrarily off the target manifold. Generating identity-preserving faces is another unsolved challenge in GAN-based face synthesis models.

In this paper, we propose a cross-domain face synthesis approach that relies on a new controllable GAN (C-GAN). It extends the original GAN by using an additional adversarial game as the third player to the GAN, competing with the refiner (generator) to preserve the specific attributes, and accordingly, providing control over the face generation process. As depicted in Figure I-1, C-GAN involves three main steps: (1) generating simulated face images via 3D morphable model (Blanz & Vetter (2003)) rendered under a specified pose, (2) refining the realism of the simulated face images using an unlabeled generic set to adapt synthetic face images in the source domain to appear as if drawn from the target domain, and (3) preserving the specific attributes of the simulated face images during the refinement through another adversarial network. Using C-GAN, a set of realistic synthetic facial images are generated that represent gallery stills under the target domain with high consistency, while preserving their identity and allowing to specify the pose conditions of synthetic images. The refined synthetic face images are then used to augment the reference gallery for FR with SSPP. The main contribution of this paper is a novel cross-domain face synthesis approach that integrates C-GAN to leverage an additional adversarial game as third player into the GAN model, producing highly consistent realistic face images in a controllable manner. Additionally, we show that using the images generated by C-GAN as additional design data within a Siamese network allows to improve still-to-video FR performance under unconstrained capture conditions.

For proof-of-concept experiments, the performance of the proposed and baseline face synthesis methods are evaluated using a "recognition via generation" framework[2] (Zhao & et al. (2017)) on videos from the public Chokepoint and COX-S2V datasets. In a particular implementation, we extend the reference gallery of a deep Siamese network for still-to-video FR.

3. Related Work

3.1 GANs for Realistic Face Synthesis.

Recently, Generative Adversarial Networks (GANs) proposed by Goodfellow & et al. (2014) have shown promising performance in face synthesis. Existing methods typically formulate GAN as a two-player game, where a discriminator *D* distinguishes face images from the real and synthesized domains, while a generator *G* reduces its discriminativeness by synthesizing

[2] In this framework, face images are synthesized first, and then the performance is evaluated with the augmented gallery.

a face of realistic quality. Their competition converges when the discriminator is unable to differentiate these two domains. Benefiting from GAN, FaceID-GAN is proposed which treats a classifier of face identity as the third player, competing with the generator by distinguishing the identities of the real and synthesized faces (Shen *et al.* (2018)). The major shortcoming of the GAN-based face synthesis models is that they may produce images that are inconsistent due to the weak global constraints. To reduce this gap, Shrivastava *et al.* (2017) developed SimGAN that learns a model using synthetic images as inputs instead of random noise vector. Our work draws inspiration from SimGAN specialized for face synthesis. Another issue of vanilla GAN is that it is difficult to control the output of the generator. Recently, conditional GANs have added condition information to the generative network and the discriminative network for conditional image synthesis (Lin *et al.* (2018b)). Tran *et al.* (2018) proposed DR-GAN takes a pose code in addition to random noise vector as the inputs for discriminator with the goal of generating a face of the same identity with the target pose that can fool the discriminator. In the CAPG-GAN, a couple-agent discriminator is introduced which forms a mask image to guide the generator in the learning process and provides a flexible controllable condition during inference (Hu *et al.* (2018)). The bottleneck of conditional GANs is the regression of the generator may lead to arbitrarily large errors in the output, which makes it unreliable for real-world applications (Chrysos *et al.* (2019)). This paper aims to address the above problems by augmenting the refiner of GAN with a domain-invariant feature extractor.

3.2 Domain-Invariant Representations.

Recently, there have been efforts to produce domain-invariant feature representations from a single input. One of the most popular approaches in this area is the domain-adversarial neural network which integrates a gradient reversal layer into the standard architecture to ensure a domain invariant feature representation (Ganin & Lempitsky (2014)). They introduced a domain confusion loss term to learn domain-invariant feature. Haeusser *et al.* (2017) presented a statistically domain invariant embedding by reinforcing associations between source and target data directly in embedding space. A slightly different approach is presented by Ghifary *et al.* (2016), where common feature assimilation is achieved implicitly by using a decoder to reconstruct the input source and target images. In a similar spirit, Sankaranarayanan *et al.* (2018) used a generator from the encoded features to generate samples which follow the same distribution as the source dataset.

4. Proposed Approach

In the following, the set $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_N\} \in \mathbb{R}^{d \times N}$ denote a gallery set composed of n reference still ROIs belonging to one of k different classes in the source domain, where d is the number of features representing a ROI and $N = kn$ is the total number of reference still ROIs; $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_i, \dots, \mathbf{y}_M\} \in \mathbb{R}^{d \times M}$ and $\mathbf{H} = \{\mathbf{h}^1, \dots, \mathbf{h}^i, \dots, \mathbf{h}^M\} \in \mathbb{R}^{(kp) \times M}$ denote the simulated set and the corresponding one-hot labels, where M is the number of the simulated ROIs and kp is number of 3D simulated classes (k identity class with p pose). The label associated with $\mathbf{y}_i$ is defined as $\mathbf{h}^i = \{h^i_d, h^i_p\}$, where h_d represents the label for identity and h_p for pose. $\mathbf{G} = \{\mathbf{g}_1, \dots, \mathbf{g}_i, \dots, \mathbf{g}_L\} \in \mathbb{R}^{d \times L}$ denote a generic set composed of L unlabeled video ROIs in the target domain. The objective of C-GAN model is to generate realistic face images with high consistency while specifying the pose, in particular pose h_p, shown in synthetic images and preserving the identity h_d.

Figure I-2 depicts the overall C-GAN process within the approach for cross-domain face synthesis. Our approach is divided into three stages: (1) 3D simulation, (2) training the refiner, (3) refiner inference. In the first stage, the 3D model of each reference still image is reconstructed via a 3D simulator and rendered under a specified pose. The rendered images, $\mathbf{Y}$, are then imported to the refiner (generator) to recover the information inherent in the target domain. In contrast to the vanilla GAN formulation (Goodfellow & et al. (2014)), in which the generator is conditioned only on a noise vector, our model's generator is constrained on both a noise vector (z) and simulated image.

During the training stage, the refiner is trained to produce realistic images through an adversarial game with a discriminator network, D_R. The discriminator classifies a refined image as real/fake image. The refiner is further encouraged to generate realistic images while preserving the identity and capture conditions of $\mathbf{Y}$ by augmenting the refiner with a domain invariant feature extractor (Ganin & Lempitsky (2014)). The feature extracting is applied on both input and output of the refiner and the Euclidean distance of the two features is considered as an additional loss. The feature extractor F must be invariant with respect to $\mathbf{Y}$ and $\mathbf{G}$ while including all identity and pose information. For this purpose, an additional adversarial game between another discriminator and the feature extractor is employed to train the feature extractor to be domain invariant. The second discriminator D_F takes the output of the domain-invariant feature extractor and distinguish between the features extracted from the the refined images and real images. In order to guaranty that the extracted features include all the information of identity and pose, an identity-pose classifier predicts identity and pose of the refined images while being trained simultaneously on the labeled 3D simulated images, $\mathbf{Y}$. In this way, the target domain variations are effectively transferred onto the reference still images while specifying the pose shown in synthetic images, and without loosing the consistency. The refiner in the proposed C-GAN shares ideas with methods for unsupervised domain adaptation (Ganin & Lempitsky (2014)), where labeled still images in the source domain and unlabeled video images from target domain are used to learn a domain-invariant embedding. We minimize the difference between the refined images and generic set while keeping the joint distribution information (on identity and pose). For stabilizing the training process of such dual-agent GAN model, we impose a boundary equilibrium regularization term. Once synthetic images are generated, any off-the-shelf classifier can be trained to perform the FR task.

4.1 3D Simulator

The simulated image set, $\mathbf{Y} \in \mathbb{R}^{d \times M}$, is formed by reconstructing the 3D face model of reference ROIs, $\mathbf{x}_i$, using a customized version of the 3DMM (Blanz & Vetter (2003)) in which the texture fitting of the original 3DMM is replaced with image mapping for simplicity (Mokhayeri *et al.* (2019a)). The shape model is defined as a convex combination of shape vectors of a set of examples:

$$\mathbf{s} = \bar{\mathbf{s}} + \sum_{k=1}^{m-1} \alpha_k . \hat{\mathbf{s}}_k \ , \tag{A I-1}$$

where $\hat{\mathbf{s}}_k$, $1 \leq k \leq m$ is the basis vector, m is the number of the basis vectors, and $\alpha_k \in [0,1]$ are the shape parameters. The optimization algorithm presented by Blanz & Vetter (2003) is employed to find optimal α_k^i, for each $\mathbf{x}_i$. Then, the material layer of $\mathbf{x}_i$ is extracted and projected to the 3D geometry of $\mathbf{x}_i$. Given the 3D facial shape and texture, novel poses can be rendered under various pose by adjusting the parameters of a camera model. During the rendering procedure, the 3D face is projected onto the image plane with weak perspective projection:

$$\mathbf{y}^j = f * \lambda * \mathbf{R}^j * (\bar{\mathbf{s}} + \sum_{k=1}^{m-1} \alpha_k^i . \hat{\mathbf{s}}_k) + \mathbf{t}_{2d}^j \,, \tag{A I-2}$$

where $\mathbf{y}^j$ is the j^{th} reconstructed pose of $\mathbf{x}_i$, f is the scale factor, λ is the orthographic projection matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, $\mathbf{R}^i$ is the rotation matrix constructed from rotation angles and $\mathbf{t}_{2d}^j$ is the translation vector.

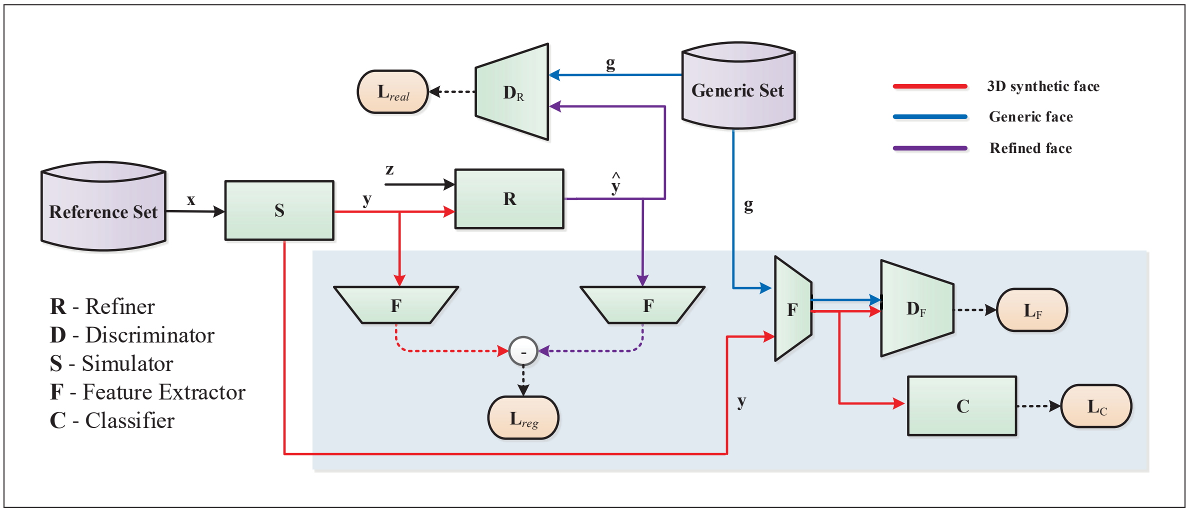


Figure-A I-2 Illustration of our proposed C-GAN architecture. It incorporates a simulator S, a refiner R, two discriminators D_R and D_F, and a classifier C constrained by the simulated and noise vector. The blue box indicates the attributes preserving module.

4.2 C-GAN Network Structure

The main part of C-GAN is the refiner (R) which improves the realism of a 3D simulator's output, $\mathbf{Y}$, using unlabeled images in the target domain, $\mathbf{G}$, while controlling their specific facial appearance (e.g. pose). The refiner, R, consists of an encoder R_{enc} and a decoder R_{dec}. The encoder R_{enc} aims to learn an identity and attribute representation from a face image $\mathbf{y}$: $\mathbf{f}(\mathbf{y}) = R_{enc}(\mathbf{y})$. The decoder R_{dec} aims to synthesize a face image $\hat{\mathbf{y}} = R_{dec}(\mathbf{f}(\mathbf{y}), \mathbf{z})$ where $\mathbf{z} \in \mathbb{R}^{N^z}$ is the noise modeling other variance besides identity or attribute (e.g. pose). The goal of R is to fool D_R to classify $\hat{\mathbf{y}}$ as a generic image.

We adopt CASIA-Net (Yi *et al.* (2014)) for R_{enc} and D_R where batch normalization and exponential linear unit are applied after each convolutional layer. A fully connected layer with logistic loss is added at the output of D_R. R_{enc} and R_{dec} are bridged by the to-be-learned identity representation $\mathbf{f}(\mathbf{y}) \in \mathbb{R}^{320}$, which is the AvgPool output in R_{enc} network. $\mathbf{f}(\mathbf{y})$ is concatenated with a random noise $\mathbf{z}$ and fed to R_{dec}. A series of fractionally-strided convolutions transforms the $(320+N_z)$-dim concatenated vector into a realistic image $\hat{\mathbf{y}} = R(\mathbf{y}, \mathbf{z})$, which is the same size as $\mathbf{y}$ (Radford *et al.* (2015)).

An encoder, F, with the same structure as R_{enc} is used at the input and output of R to compare the domain invariant features of the simulated images to the refined images. As mentioned before, the discriminator, D_F, and classifier, C, are used to train F. Both C and D_F consist two fully-connected 1024 unit layers. A fully-connected softmax loss for k identity and p pose classification is added to C while a 1 unit fully-connected logistic layer is added to D_F. Table I-1 shows the neural network structures of R_{enc}, D_R, F and R_{enc}.

4.3 Training the Refiner

Let a refined image be denoted by $\hat{\mathbf{y}}_i$, then $\hat{\mathbf{y}} = R(\theta_R; \mathbf{y})$ where θ_R is the function parameters. The key requirement is that the refined image, $\hat{\mathbf{y}}$, must look like a real generic image preserving the identity and pose information from the 3D simulator. To this end, θ_R is learned by minimizing a combination of two losses:

$$\mathcal{L}_R(\theta_R, \theta_F) = \sum_i \mathcal{L}_{real}(\theta_R; \mathbf{y}_i, \mathbf{G}) + \lambda \mathcal{L}_{reg}(\theta_F; \mathbf{y}_i) \tag{A I-3}$$

Table-A I-1 The network structure of the proposed C-GAN architecture.

R_{end} and D_R			R_{dec}		
Layer	Filter/Stride	Output Size	Layer	Filter/Stride	Output Size
			FConv52	$3 \times 3/1$	$6 \times 6 \times 320$
Conv11	$3 \times 3/1$	$96 \times 96 \times 32$	FConv52	$3 \times 3/1$	$6 \times 6 \times 160$
Conv12	$3 \times 3/1$	$96 \times 96 \times 64$	FConv51	$3 \times 3/1$	$6 \times 6 \times 256$
Conv21	$3 \times 3/2$	$48 \times 48 \times 64$	FConv43	$3 \times 3/2$	$12 \times 12 \times 256$
Conv22	$3 \times 3/1$	$48 \times 48 \times 64$	FConv42	$3 \times 3/1$	$12 \times 12 \times 128$
Conv23	$3 \times 3/1$	$48 \times 48 \times 128$	FConv41	$3 \times 3/1$	$12 \times 12 \times 192$
Conv31	$3 \times 3/2$	$24 \times 24 \times 128$	FConv33	$3 \times 3/2$	$24 \times 24 \times 192$
Conv32	$3 \times 3/1$	$24 \times 24 \times 96$	FConv32	$3 \times 3/1$	$24 \times 24 \times 96$
Conv33	$3 \times 3/1$	$24 \times 24 \times 192$	FConv31	$3 \times 3/1$	$24 \times 24 \times 128$
Conv41	$3 \times 3/2$	$12 \times 12 \times 192$	FConv23	$3 \times 3/2$	$48 \times 48 \times 128$
Conv42	$3 \times 3/1$	$12 \times 12 \times 128$	FConv22	$3 \times 3/1$	$48 \times 48 \times 64$
Conv43	$3 \times 3/1$	$12 \times 12 \times 256$	FConv21	$3 \times 3/1$	$48 \times 48 \times 64$
Conv51	$3 \times 3/2$	$6 \times 6 \times 256$	FConv13	$3 \times 3/2$	$96 \times 96 \times 64$
Conv52	$3 \times 3/1$	$6 \times 6 \times 160$	FConv12	$3 \times 3/1$	$96 \times 96 \times 32$
Conv53	$3 \times 3/1$	$6 \times 6 \times 320$	FConv11	$3 \times 3/1$	$96 \times 96 \times 1$
AvgPool	$6 \times 6/1$	$1 \times 1 \times 320$			
FC (D_R only)		1			

The first part of the cost, $\mathcal{L}_{real}$, adds realism to the simulated images, while the second part, $\mathcal{L}_{reg}$, preserves the identity and pose information.

The adversarial loss used for training the refiner network, R, is responsible for fooling D_R into classifying the refined images as real. This problem is modeled a two-player minimax game, and update the refiner network, R, and the discriminator network. D_R updates its parameters by minimizing the following loss:

$$\mathcal{L}_D(\phi) = -\sum_i \log(D_R(\phi; \hat{\mathbf{y}}_i)) - \sum_j \log(1 - D_R(\phi; \mathbf{g}_j)) \tag{A I-4}$$

where $D_R(\cdot)$ is the probability of the input being a refined image, and $1 - D_R(\cdot)$ that of a real one. For training this network, each mini-batch consists of randomly sampled $\hat{\mathbf{y}}_\mathbf{i}$ and $\mathbf{g}_\mathbf{i}$.

The realism loss function employs the trained discriminator D_R as follows:

$$\mathcal{L}_{real}(\theta_R) = \log(1 - D_R(R(\theta_R; \mathbf{y}_i))) \tag{A I-5}$$

By minimizing this loss function, the refiner forces the discriminator to fail classifying the refined images as synthetic.

In order to preserve the annotation information of the 3D simulator, we use a self-regularization loss that minimizes difference between a feature transform of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$,

$$\mathcal{L}_{reg}(\theta_F) = \|F(\hat{\mathbf{y}}_i, \theta_F) - F(\mathbf{y}_i, \theta_F)\| \tag{A I-6}$$

where F is the mapping from image space to a feature space, and $\|\cdot\|$ is the ℓ_2 norm.

Another adversarial game is employed to train the feature extractor network parameters (θ_F). For this purpose, the classifier, $C(.)$, assigns identity and pose information labels ($\mathbf{h}^i$) to a set of features extracted by F. In this way, F learns to extract the features that are domain-invariant and consist information of identity and pose. F and C are updated based on the identity and pose labels of $\mathbf{Y}$ in a traditional supervised manner. F is also updated using the adversarial gradients from D_F so that the feature learning and image generation processes co-occur smoothly.

$$\mathcal{L}_C(\theta_C) = -\sum_i \sum_{j=1}^{c} \mathbf{h}_j^i \log(C(\theta_C; F(\hat{\mathbf{y}}_i))) \tag{A I-7}$$

$$\mathcal{L}_{D_F}(\gamma) = -\sum_i \log(D_F(\gamma; F(\hat{\mathbf{y}}_i))) - \sum_i \log(1 - D_F(\gamma; F(\mathbf{g}_j))) \tag{A I-8}$$

Given a realistic simulated images $\hat{\mathbf{y}}_i$ as input, D_F outputs a binary distribution optimized by minimizing a binary cross entropy loss $\mathcal{L}_F$.The gradients are generated using the following loss functions:

$$\mathcal{L}_F(\theta_F) = \sum_i \log(1 - D_F(F(\theta_F \hat{\mathbf{y}}_i))) \tag{A I-9}$$

where the F and D_F parameters are learned by minimizing $\mathcal{L}_F(\theta_F)$ and $\mathcal{L}_{D_F}(\gamma)$ alternately. We leave γ fixed while updating the parameters of F, and we fix θ_F while updating D_F.

5. Experimental Analysis

5.1 Evaluation Methodology

The performance of the proposed and baseline methods was evaluated using two datasets for still-to-video FR. The **Chokepoint** (Wong *et al.* (2011)) consists of 25 subjects walking through portal 1 and 29 subjects in portal 2. Videos are recorded over 4 sessions one month apart. An array of 3 cameras are mounted above portal 1 and portal 2 that capture the entry of subjects during 4 sessions. In total, the dataset consists of 54 video sequences and $64,204$ face images. **COX-S2V** dataset (Huang *et al.* (2015)) contains 1000 individuals, with 1 high-quality still image and 4 low-resolution video sequences per individual simulating video surveillance scenario. The video frames are captured by 4 cameras mounted at fixed locations. In each video, an individual walks through a designed S-shape route with changes in illumination, scale, and pose.

FR performance under SSPP scenario was assessed via the "recognition via generation" framework to validate our hypothesis that adding photo-realistic synthetic reference faces to the gallery set can address the visual domain shift, and accordingly improve the accuracy.. Besides, since photographic results also indicate the performances qualitatively, the visual quality is also compared in our experiment. We also compared our results with those obtained by flow-based Frontalization (Hassner *et al.* (2015)). During the enrollment of an individual to the system, q simulated ROIs for each reference still ROI are generated under different poses using the conventional 3DMM (Blanz & Vetter (2003)). The images are then refined using the controlled GAN that projects the capture conditions of the target domain on them while preserving their pose and identity. The gallery is formed using the original reference still ROIs along with the corresponding synthetic ROIs. During the operational phase, FR is performed using Siamese network model that is pre-trained using the VGG-Face2 dataset with Inception Resnet V1 architecture. The CNN feature extractors in this model is trained using stochastic gradient descent and AdaGrad with standard back-propagation (Schroff *et al.* (2015)). Finally, given the query (video) and reference (still) feature vectors, pair-wise matching is performed using the k-NN classifier based on Euclidean distance.

In all experiments with Chokepoint and COX-S2V datasets, 5 and 20 target individuals are selected, respectively, to populate the watch-list, using one high-quality still image. Videos of 10 individuals that are assumed to come from unlabeled persons are used as a generic set. The rest of the videos including 10 other unlabeled individuals and 5 videos of the individuals who are already enrolled in the watch-list are used for testing. In order to obtain representative results, this process is repeated 5 times with a different random selection of watch-lists and the average performance is reported with standard deviation over all the runs. The average performance of the proposed and baseline system for still-to-video FR is presented by measuring the partial area under ROC curve, pAUC(20%) (using the AUC at $0 < FPR \leq 20\%$), and mean average precision, mAP. We further employed the Frechet Inception Distance (FID) (Heusel *et al.* (2017)) to quantitatively verify the superiority of our synthetic faces.

5.2 Results and Discussion

Figure I-3 shows examples of the synthetic face images generated based on our proposed technique using the original reference still ROIs and generic set of the Chokepoint dataset. This show that the images refined using C-GAN can preserve their pose variations. Figure I-4 compares the qualitative results obtained with state-of-the-art techniques; (b) 3DMM (Blanz & Vetter (2003)), (c) 3DMM-CNN (Tran *et al.* (2017a)), (d) DSFS (Mokhayeri *et al.* (2019a)), and (e) our proposed C-GAN.

Table I-2 shows the average accuracy of a deep Siamese network for still-to-video FR that relies on the proposed and baseline methods for generating synthetic face images to augment the reference gallery. The baseline system is designed with an original reference still ROI alone. For our proposed C-GAN technique, the synthetic faces are generated with 5° step size within a range of ±5 to ±60 degrees in yaw, pitch, and roll. Consequently, we have $q = 73$ synthetic face images in total for our experiments. For reference, the still-to-video FR system based on frontalization is also evaluated. Results indicate that by adding extra synthetic ROIs generated with C-GAN allows to outperform baseline systems. pAUC and mAP accuracy increases by about 3%, typically with $q = 73$ synthetic pose ROIs for Chokepoint and COX-

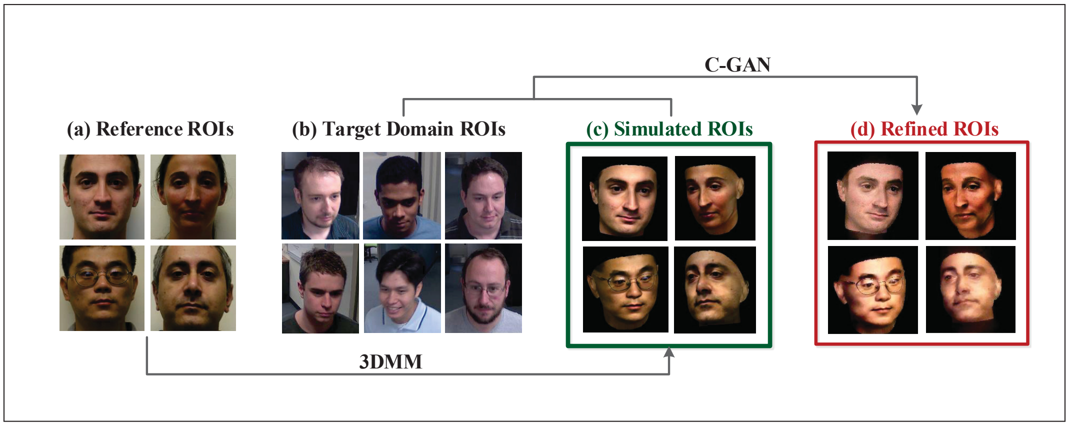


Figure-A I-3 Examples of the synthetic faces obtained with the proposed approach on Chokepoint database (ID#1, ID#5, ID#6, ID#16). The simulated images (c) are refined based on the target domain capture conditions (b) while preserving the identity of reference stills (a) under specific pose.

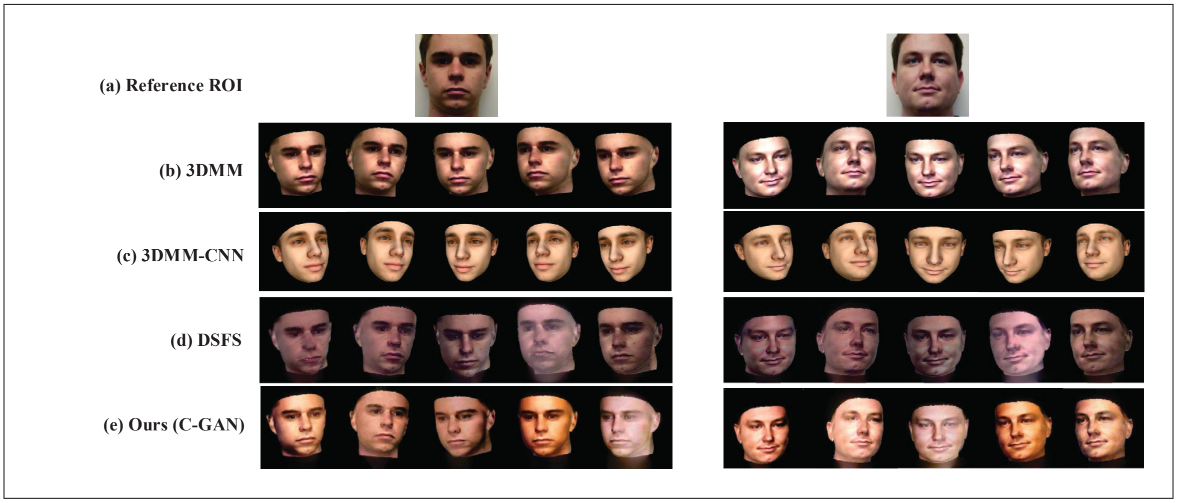


Figure-A I-4 Examples of facial images generated using state-of-the-art face synthesizing methods on Chokepoint dataset (ID#23, ID#25).

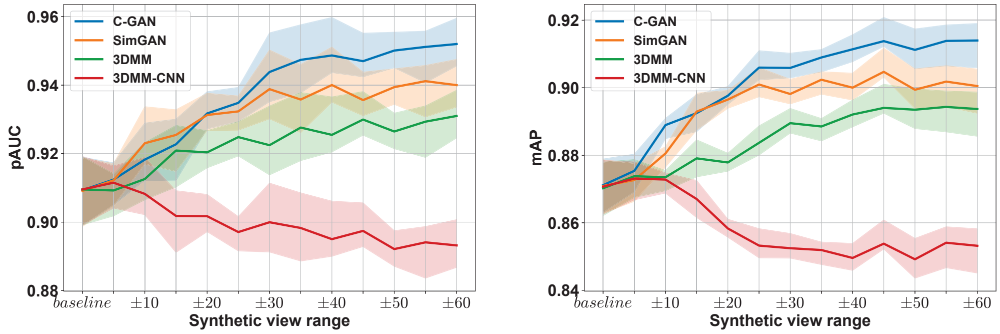
S2V datasets. Results suggest that that leveraging target domain information within the GAN framework while controlling its pose and identity can efficiently mitigate the impacts of the visual domain shift.

Figure I-5 shows the average pAUC(20%) (a) and mAP (b) accuracy obtained for the implementation of still-to-video FR when increasing the number of synthetic ROIs per each individ-

Table-A I-2 Average pAUC and mAP accuracy of the Siamese network using the proposed and baseline methods for data augmentation. The '# synth' columns show the minimum number of synthetic samples needed to attain the highest level of accuracy.

Techniques	**Chokepoint database**			**COX-S2V database**		
	pAUC(20%)	**mAP**	**# Synth**	**pAUC**(20%)	**mAP**	**# Synth**
Baseline	0.908±0.018	0.861±0.020	N/A	0.912±0.017	865±0.016	N/A
3DMM (Blanz & Vetter (2003))	0.917±0.023	0.877±0.025	73	0.928±0.026	872±0.027	73
3DMM-CNN (Tran *et al.* (2017a))	0.915±0.025	0.873±0.028	73	0.922±0.024	871±0.028	73
DSFS (Mokhayeri *et al.* (2019a))	0.923±0.018	0.880±0.019	17	0.934±0.021	896±0.022	14
SimGAN (Shrivastava *et al.* (2017))	0.942±0.025	0.901±0.023	73	0.948±0.023	904±0.020	73
DR-GAN (Tran *et al.* (2018))	0.931±0.019	0.893±0.017	73	0.939±0.016	903±0.017	73
FaceID-GAN (Shen *et al.* (2018))	0.936±0.023	0.905±0.019	73	0.942±0.018	911±0.022	73
Dual-GAN (Zhao & et al. (2017))	0.948±0.021	0.915±0.018	73	0.952±0.021	922±0.024	73
Frontalization (Hassner *et al.* (2015))	0.919±0.020	0.884±0.019	N/A	0.926±0.017	892±0.020	N/A
C-GAN (Ours)	0.951±0.023	0.917±0.022	73	0.957±0.019	925±0.019	73

ual. Adding synthetic ROIs generated under various capture conditions allows to significantly outperform the baseline system designed with the original reference still ROI alone. As shown in I-5, accuracy trends to stabilize to its maximum value when the size of the synthetic faces is greater than $q = 73$ in C-GAN.



(a) Chokepoint (b) COX-S2V

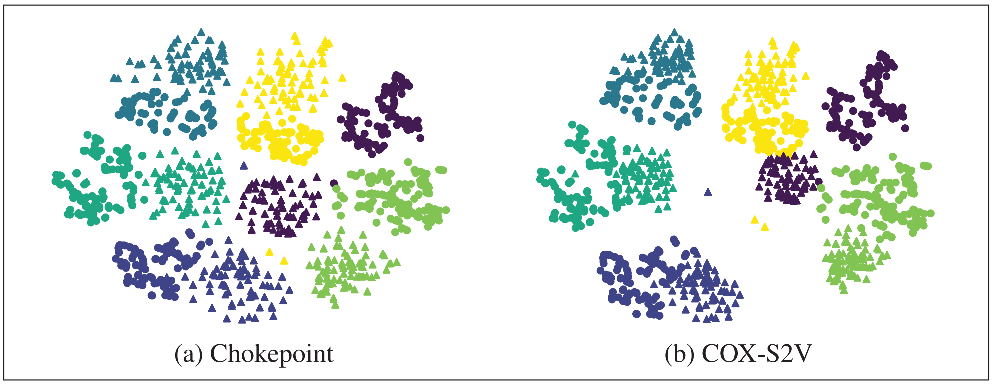
Figure-A I-5 Average pAUC(20%) (a) and mAP (b) accuracy of the proposed and baseline techniques versus the size of the synthetic set on Chokepoint database.

Frechet Inception Distance (FID) (Heusel *et al.* (2017)) has been recently proposed to evaluate the performance of image synthesis tasks quantitatively where lower FID score indicates the smaller Wasserstein distance between two distributions. Inception V3 model is employed to extract feature vectors from images. Table I-3 show the FID between the real and the synthesized faces across different yaw which demonstrates the effectiveness of our method.

Table-A I-3 Frechet Inception Distance (FID) across different views with Chokepoint and COX-S2V datasets.

Technique	**Chokepoint data**			**COX-S2V data**		
	±5°	±15°	±45°	±5°	±15°	±45°
3DMM (Blanz & Vetter (2003))	22.3	23.4	25.7	20.5	21.4	21.7
3DMM-CNN (Tran *et al.* (2017a))	49.5	53.2	61.4	42.2	50.7	53.2
DSFS (Mokhayeri *et al.* (2019a))	21.4	22.7	24.5	17.9	21.8	23.1
C-GAN (Ours)	20.9	22.1	23.8	17.3	20.9	21.5

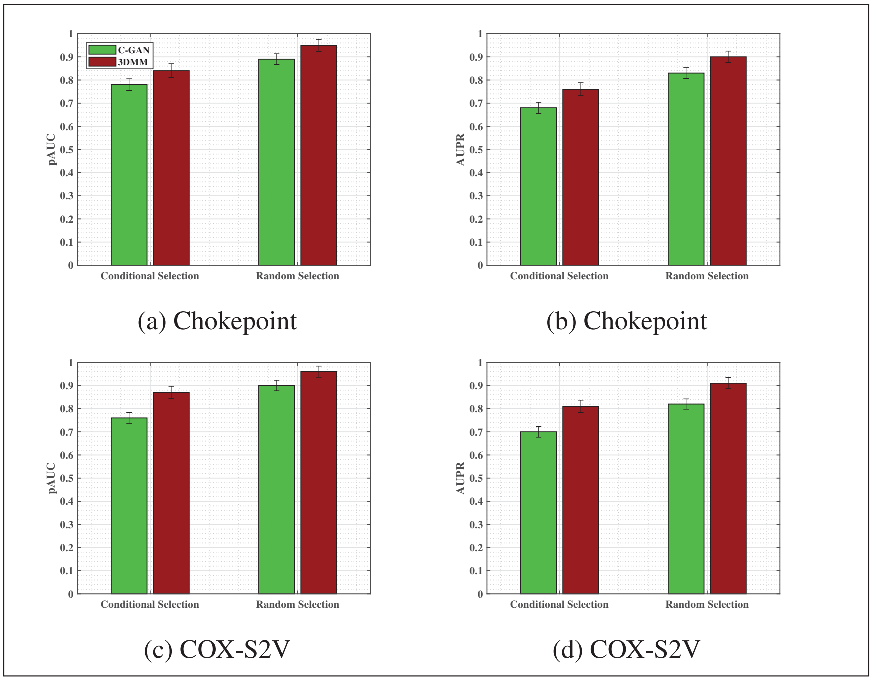
To further evaluate the effectiveness of refiner in our C-GAN, we use t-SNE (Maaten & Hinton (2008)) to visualize the deep features of simulated, refined and real faces in a 2D space. Figure I-6 shows there is significant difference between the distribution of 3D simulated and real face. However, after refining the 3D simulated images, the distribution of the refined images become closer to the distribution of the real images.


(a) Chokepoint

(b) COX-S2V

Figure-A I-6 t-SNE visualization. Circles represent the generic set. Triangles in (a) represent 3D simulated faces while triangles in (b) represent refined faces.

Figure I-7 compares the performance of Siamese networks for FR when adding 73 selected synthetic ROIs generated with the C-GAN versus 73 randomly selected images (without condition). For reference, FR based on 3DMM face synthesizing is also evaluated. Results in this figure show that the C-GAN with a specified range outperforms other models – FR performance is higher when the gallery is designed using the representative views than based gallery comprised of randomly selected synthetic faces per person. The proposed C-GAN can therefore adequately generate representative facial ROIs for the reference gallery.



(a) Chokepoint

(b) Chokepoint

(c) COX-S2V

(d) COX-S2V

Figure-A I-7 Average pAUC(20%) and AUPR accuracy for Siamese network with C-GAN and 3DMM face synthesis with q specified and randomly selected faces on Chokepoint (a,b) and COX-S2V (c,d) databases. Error bars are standard deviation.

5.2.1 Ablation Study

To evaluate the components of C-GAN (D_F , D_R , C), the model is trained by removing these modules while fixing the training process and all parameters. Recognition accuracy is evaluated on the synthetic images generated from each variant. We observe (Table I-4) that the accuracy will decrease by about 3% if one module is not used.

Table-A I-4 Results of ablation study with Chokepoint and COX-S2V datasets.

Accuracy	**Removed Module**					
	Chokepoint data			**COX-S2V database**		
	D_F	D_R	C	D_F	D_R	**C**
pAUC	0.905±0.022	0.901±0.023	0.882±0.020	0.908±0.021	0.902±0.025	0.891±0.027
mAP	0.873±0.024	0.868±0.021	0.854±0.019	0.885±0.018	0.875±0.020	0.859±0.024

5.2.2 Time Complexity

Time complexity is estimated empirically, using the amount of time required to match 2 facial ROIs with the given dataset. The average running time is measured with a randomly selected probe ROIs using a PC workstation with an Intel Core i7 CPU (3.41GHz) processor and 16GB RAM. Table I-5 shows average matching time of deep Siamese networks over videos ROIs of the Chokepoint and COX-S2V datasets. The table shows time complexity grows the gallery size. The results suggest that the proposed approach represents an interesting trade-off between accuracy and complexity.

Table-A I-5 Average matching time over videos ROIs of the Chokepoint and COX-S2V datasets.

Techniques	**Matching Time (sec)**	
	Chokepoint	**COX-S2V**
Siamese Network (Koch *et al.* (2015))		
· 1 frontal reference still / person	6.4	13.2
· +73 uniform synthetic / person	129.7	186.1
· +100 random synthetic / person	152.3	211.5
Frontalization (Hassner *et al.* (2015))	12.7	16.3

6. Conclusion

In this paper, a cross-domain face synthesis approach with a new C-GAN model is proposed for data augmentation that generates highly consistent, realistic and identity preserving synthetic face images under specific pose conditions. The proposed model allows to mitigate the impact of some common issues with the original GAN model for data augmentation, such as lack of control and inconsistency. C-GAN leverages an additional adversarial game as third player to encourage the refiner during the inference to specify the capture conditions shown in synthetic images in a controllable manner. This allows augmenting to the gallery of a deep Siamese network with a diverse, yet compact set of synthetic views relevant to the target domain. Experimental results obtained using the Chokepoint and COX-S2V datasets suggest that the synthetic face images based on C-GAN allow us address visual domain shift, and thereby improve the accuracy of still-to-video FR system, with no need to generate a large number of synthetic face images. A future direction is to simulate and control other facial appearance (e.g. illumination and expression) during the face synthesis process. This can be further used to augment a dataset with representative images to train a deep neural network for FR systems.

BIBLIOGRAPHY

Aharon, M., Elad, M. & Bruckstein, A. (2006). k-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. on IP*, 54(11), 4311–4322.

Ahmed, E., Jones, M. & Marks, T. K. (2015). An improved deep learning architecture for person re-identification. *CVPR*.

Almageed, W., Wu, Y., Rawls, S., Harel, S., Hassner, T., Masi, I., Choi, J., Lekust, J., Kim, J., Natarajan, P. et al. (2016). Face recognition using deep multi-pose representations. *WACV*.

Arjovsky, M., Chintala, S. & Bottou, L. (2017). Wasserstein generative adversarial networks. *ICML*.

Asthana, A., Zafeiriou, S., Cheng, S. & Pantic, M. (2013). Robust discriminative response map fitting with constrained local models. *CVPR*.

Baltruvsaitis, T., Robinson, P. & Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. *WACV*.

Bao, J., Chen, D., Wen, F., Li, H. & Hua, G. (2018a). Towards open-set identity preserving face synthesis. *CVPR*.

Bao, J., Chen, D., Wen, F., Li, H. & Hua, G. (2018b). Towards open-set identity preserving face synthesis. *CVPR*.

Bashbaghi, S., Granger, E., Sabourin, R. & Bilodeau, G. A. (2014). Watch-List Screening Using Ensembles Based on Multiple Face Representations. *ICPR*.

Bashbaghi, S., Granger, E., Sabourin, R. & Bilodeau, G.-A. (2015). Ensembles of exemplar-SVMs for video face recognition from a single sample per person. *AVSS*.

Bashbaghi, S., Granger, E., Sabourin, R. & Bilodeau, G.-A. (2017a). Dynamic ensembles of exemplar-SVMs for still-to-video face recognition. *Pattern Recognition*, 69, 61–81.

Bashbaghi, S., Granger, E., Sabourin, R. & Bilodeau, G.-A. (2017b). Robust watch-list screening using dynamic ensembles of SVMs based on multiple face representations. *Machine Vision and Applications*, 28, 219–241.

Berthelot, D., Schumm, T. & Metz, L. (2017). Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*.

Blanz, V. & Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. *IEEE Trans. on PAMI*, 25(9), 1063–1074.

Brock, A., Donahue, J. & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. *ICLR*.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E. & Shah, R. (1994). Signature verification using a" siamese" time delay neural network. *NIPS*.

Cai, S., Zhang, L., Zuo, W. & Feng, X. (2016). A probabilistic collaborative representation based approach for pattern classification. *CVPR*.

Cao, J., Hu, Y., Yu, B., He, R. & Sun, Z. (2019a). 3D Aided Duet GANs for Multi-view Face Image Synthesis. *IEEE Trans. on IFS*.

Cao, J., Hu, Y., Zhang, H., He, R. & Sun, Z. (2019b). Towards High Fidelity Face Frontalization in the Wild. *International Journal of Computer Vision*, 1–20.

Cao, K., Rong, Y., Li, C., Tang, X. & Change Loy, C. (2018). Pose-robust face recognition via deep residual equivariant mapping. *CVPR*.

Cevikalp, H., Yavuz, H. S. & Triggs, B. (2019). Face Recognition Based on Videos by Using Convex Hulls. *IEEE Trans. on Circuits and Systems for Video Technology*.

Chen, C. & Ross, A. (2019). Matching Thermal to Visible Face Images Using a Semantic-Guided Generative Adversarial Network. *FG*.

Chen, Q. & Koltun, V. (2017). Photographic image synthesis with cascaded refinement networks. *ICCV*.

Chen, X., Wu, H., Jin, X. & Zhao, Q. (2013). Face illumination manipulation using a single reference image by adaptive layer decomposition. *IEEE Trans. on IP*, 22(11), 4249–4259.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S. & Choo, J. (2018). StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. *CVPR*.

Chrysos, G., Kossaifi, J. & Zafeiriou, S. (2019). Robust Conditional Generative Adversarial Networks. *ICLR*.

De-la Torre, M., Granger, E., Sabourin, R. & Gorodnichy, D. O. (2015). Adaptive skew-sensitive ensembles for face recognition in video surveillance. *Pattern Recognition*, 48(11), 3385–3406.

Deng, J., Guo, J., Xue, N. & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. *CVPR*.

Deng, W., Yin, W. & Zhang, Y. (2013). Group sparse optimization by alternating direction method. *SPIE*, 8858, 88580R.

Deng, W., Hu, J. & Guo, J. (2012). Extended SRC: Undersampled face recognition via intraclass variant dictionary. *IEEE Trans. on PAMI*, 34(9), 1864–1870.

Deng, W., Hu, J., Zhou, X. & Guo, J. (2014). Equidistant prototypes embedding for single sample based face recognition with generic learning and incremental learning. *Pattern Recognition*, 47(12), 3738–3749.

Deng, W., Hu, J. & Guo, J. (2018). Face recognition via collaborative representation: its discriminant nature and superposed representation. *IEEE Trans. on PAMI*, 40(10), 2513–2521.

Dewan, M. A. A., Granger, E., Marcialis, G.-L., Sabourin, R. & Roli, F. (2016). Adaptive appearance model tracking for still-to-video face recognition. *Pattern Recognition*, 49, 129–151.

Elhamifar, E. & Kaluza, M. C. D. P. (2017). Online Summarization via Submodular and Convex Optimization. *CVPR*.

Elhamifar, E. & Vidal, R. (2011). Robust classification using structured sparse representation. *CVPR*.

Elhamifar, E., Sapiro, G. & Vidal, R. (2012). Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. *NIPS*.

Fan, Z., Zhang, D., Wang, X., Zhu, Q. & Wang, Y. (2018). Virtual dictionary based kernel sparse representation for face recognition. *Pattern Recognition*, 76, 1–13.

Frey, B. J. & Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814), 972–976.

Ganin, Y. & Lempitsky, V. (2014). Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*.

Gao, S., Jia, K., Zhuang, L. & Ma, Y. (2015). Neither global nor local: Regularized patch-based representation for single sample per person face recognition. *International Journal of Computer Vision*, 111(3), 365–383.

Gao, Y., Ma, J. & Yuille, A. L. (2017). Semi-Supervised Sparse Representation Based Classification for Face Recognition With Insufficient Labeled Samples. *IEEE Trans. on IP*, 26(5), 2545–2560.

Gecer, B., Bhattarai, B., Kittler, J. & Kim, T.-K. (2018). Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3D morphable model. *ECCV*.

Gecer, B., Ploumpis, S., Kotsia, I. & Zafeiriou, S. (2019). GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction. *CVPR*.

Ghifary, M., Kleijn, W., Zhang, M., Balduzzi, D. & Li, W. (2016). Deep reconstruction-classification networks for unsupervised domain adaptation. *ECCV*.

Gonzalez-Garcia, A., van de Weijer, J. & Bengio, Y. (2018). Image-to-image translation for cross-domain disentanglement. *NIPS*.

Goodfellow, I. & et al. (2014). Generative adversarial nets. *NIPS*.

Haeusser, P., Frerix, T., Mordvintsev, A. & Cremers, D. (2017). Associative domain adaptation. *ICCV*.

Hassner, T., Harel, S., Paz, E. & Enbar, R. (2015). Effective face frontalization in unconstrained images. *CVPR*.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *CVPR*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*.

Hong, S., Im, W., Ryu, J. & Yang, H. S. (2017). SSPP-DAN: Deep domain adaptation network for face recognition with single sample per person. *ICIP*.

Hu, Y., Wu, X., Yu, B., He, R. & Sun, Z. (2018). Pose-guided photorealistic face rotation. *CVPR*.

Huang, G. B., Mattar, M., Berg, T. & Learned-Miller, E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments.

Huang, G. B., Mattar, M., Lee, H. & Learned-Miller, E. (2012). Learning to Align from Scratch. *NIPS*.

Huang, R., Zhang, S., Li, T., He, R. et al. (2017a). Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. *ICCV*.

Huang, Z., Shan, S., Wang, R., Zhang, H., Lao, S., Kuerban, A. & Chen, X. (2015). A benchmark and comparative study of video-based face recognition on COX face database. *IEEE Trans. on IP*, 24(12), 5967–5981.

Huang, Z., Wang, R., Shan, S., Van Gool, L. & Chen, X. (2017b). Cross euclidean-to-riemannian metric learning with application to face recognition from video. *IEEE trans. on PAMI*, 40(12), 2827–2840.

Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *CVPR*.

Jeon, J., Cho, S., Tong, X. & Lee, S. (2014). Intrinsic image decomposition using structure-texture separation and surface normals. *ECCV*.

Ji, H.-K., Sun, Q.-S., Ji, Z.-X., Yuan, Y.-H. & Zhang, G.-Q. (2017). Collaborative probabilistic labels for face recognition from single sample per person. *Pattern Recognition*, 62, 125–134.

Jiang, L., Zhang, J. & Deng, B. (2019). Robust RGB-D Face Recognition Using Attribute-Aware Loss. *IEEE Trans. on PAMI*.

Jiao, L., Zhang, S., Li, L., Liu, F. & Ma, W. (2018). A modified convolutional neural network for face sketch synthesis. *Pattern Recognition*, 76, 125–136.

Kazemi, V. & Josephine, S. (2014). One millisecond face alignment with an ensemble of regression trees. *CVPR*.

Koch, G., Zemel, R. & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. *ICML workshop*.

Koppen, P., Feng, Z.-H., Kittler, J., Awais, M., Christmas, W., Wu, X.-J. & Yin, H.-F. (2018). Gaussian mixture 3D morphable face model. *Pattern Recognition*, 74, 617–628.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *NIPS*.

Li, P., Prieto, L., Mery, D. & Flynn, P. (2018). Face Recognition in Low Quality Images: A Survey. *arXiv preprint arXiv:1805.11519*.

Li, Z.-M., Huang, Z.-H. & Shang, K. (2016). A customized sparse representation model with mixed norm for undersampled face recognition. *IEEE Trans. on IFS*, 11(10), 2203–2214.

Lin, G., Yang, M., Yang, J., Shen, L. & Xie, W. (2018a). Robust, discriminative and comprehensive dictionary learning for face recognition. *Pattern Recognition*, 81, 341–356.

Lin, J., Xia, Y., Qin, T., Chen, Z. & Liu, T. (2018b). Conditional image-to-image translation. *CVPR*.

Liu, F., Tran, L. & Liu, X. (2019). 3D Face Modeling from Diverse Raw Scan Data. *ICCV*.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B. & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. *CVPR*.

Liu, Y. & Wassell, I. J. (2015). A New Face Recognition Algorithm based on Dictionary Learning for a Single Sample per Person. *BMVC*.

Lu, J., Tan, Y.-P. & Wang, G. (2013). Discriminative multimanifold analysis for face recognition from a single training sample per person. *IEEE Trans. on PAMI*, 35(1), 39–51.

Luo, X., Xu, Y. & Yang, J. (2019). Multi-resolution dictionary learning for face recognition. *Pattern Recognition*, 93, 283–292.

Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9, 2579–2605.

Masi, I., Tran, A. T., Hassner, T., Leksut, J. T. & Medioni, G. (2016). Do we really need to collect millions of faces for effective face recognition? *ECCV*.

Masi, I., Chang, F.-J., Choi, J., Harel, S., Kim, J., Kim, K., Leksut, J., Rawls, S., Wu, Y., Hassner, T. et al. (2019a). Learning pose-aware models for pose-invariant Face Recognition in the wild. *IEEE Trans. on PAMI*, 41(2), 379–393.

Masi, I., Tran, A. T., Hassner, T., Sahin, G. & Medioni, G. (2019b). Face-Specific Data Augmentation for Unconstrained Face Recognition. *International Journal of Computer Vision*, 1–26.

Migneault, F. C., Granger, E. & Mokhayeri, F. (2018). Using Adaptive Trackers for Video Face Recognition from a Single Sample Per Person. *IPTA*.

Mokhayeri, F., Granger, E. & Bilodeau, G.-A. (2019a). Domain-specific face synthesis for video face recognition from a single sample per person. *IEEE Trans. on IFS*, 14(3), 757–772.

Mokhayeri, F. & Granger, E. (2018). Robust Video Face Recognition From a Single Still Using a Synthetic Plus Variational Model. *FG*.

Mokhayeri, F., Granger, E. & Bilodeau, G.-A. (2015). Synthetic face generation under various operational conditions in video surveillance. *ICIP*.

Mokhayeri, F., Kamali, K. & Granger, E. (2019b). Cross-Domain Face Synthesis using a Controllable GAN. *arXiv preprint arXiv:1910.14247*.

Ni, J., Qiu, Q. & Chellappa, R. (2013). Subspace interpolation via dictionary learning for unsupervised domain adaptation. *CVPR*.

Nourbakhsh, F., Granger, E. & Fumera, G. (2016). An extended sparse classification framework for domain adaptation in video surveillance. *ACCV*.

Pagano, C., Granger, E., Sabourin, R., Marcialis, G. L. & Roli, F. (2014). Adaptive ensembles for face recognition in changing video surveillance environments. *Information Sciences*, 286, 75–101.

Parchami, M., Bashbaghi, S. & Granger, E. (2017a). CNNs with cross-correlation matching for face recognition in video surveillance using a single training sample per person. *AVSS*.

Parchami, M., Bashbaghi, S. & Granger, E. (2017b). Video-based face recognition using ensemble of haar-like deep convolutional neural networks. *IJCNN*.

Parchami, M., Bashbaghi, S., Granger, E. & Sayed, S. (2017c). Using deep autoencoders to learn robust domain-invariant representations for still-to-video face recognition. *AVSS*.

Parkhi, O. M., Vedaldi, A. & Zisserman, A. (2015). Deep Face Recognition. *BMVC*.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S. & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. *AVSS*.

Peng, C., Wang, N., Li, J. & Gao, X. (2019). DLFace: Deep local descriptor for cross-modality Face Recognition. *Pattern Recognition*, 90, 161–171.

Radford, A., Metz, L. & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rakotomamonjy, A. (2011). Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms. *Signal Processing*, 91(7), 1505–1526.

Richardson, E., Sela, M., Or-El, R. & Kimmel, R. (2017). Learning detailed face reconstruction from a single image. *CVPR*.

Salakhutdinov, R. & Hinton, G. (2007). Learning a nonlinear embedding by preserving class neighbourhood structure. *Artificial Intelligence and Statistics*.

Sankaranarayanan, S., Balaji, Y., Castillo, C. & Chellappa, R. (2018). Generate to adapt: Aligning domains using generative adversarial networks. *CVPR*.

Sanyal, S., Bolkart, T., Feng, H. & Black, M. (2019). Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision. *CVPR*.

Schroff, F., Kalenichenko, D. & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *CVPR*.

Shao, C., Song, X., Feng, Z.-H., Wu, X.-J. & Zheng, Y. (2017). Dynamic dictionary optimization for sparse-representation-based face classification using local difference images. *Information Sciences*, 393, 1–14.

Shekhar, S., Patel, V. M. & Chellappa, R. (2017). Synthesis-based robust low resolution face recognition. *arXiv preprint arXiv:1707.02733*.

Shen, Y., Luo, P., Yan, J., Wang, X. & Tang, X. (2018). FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis. *CVPR*.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W. & Webb, R. (2017). Learning from Simulated and Unsupervised Images through Adversarial Training. *CVPR*.

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ICLR*.

Sohn, K., Liu, S., Zhong, G., Yu, X., Yang, M. & Chandraker, M. (2017). Unsupervised domain adaptation for face recognition in unlabeled videos. *CVPR*.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *NIPS*.

Su, Y., Shan, S., Chen, X. & Gao, W. (2010). Adaptive generic learning for face recognition from a single sample per person. *CVPR*.

Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. *CVPR*.

Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Pérez, P. & Theobalt, C. (2017). MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. *ICCV*.

Tran, A. T., Hassner, T., Masi, I. & Medioni, G. (2017a). Regressing robust and discriminative 3D morphable models with a very deep neural network. *CVPR*.

Tran, L. Q., Yin, X. & Liu, X. (2018). Representation learning by rotating your faces. *IEEE Trans. on PAMI*.

Tran, L. & Liu, X. (2018). Nonlinear 3D face morphable model. *CVPR*, pp. 7346–7355.

Tran, L. & Liu, X. (2019). On learning 3D face morphable model from in-the-wild images. *IEEE Trans. on PAMI*.

Tran, L., Yin, X. & Liu, X. (2017b). Disentangled representation learning GAN for pose-invariant face recognition. *CVPR*.

Tran, L., Liu, F. & Liu, X. (2019). Towards high-fidelity nonlinear 3D face morphable model. *CVPR*.

Tropp, J. A., Gilbert, A. C. & Strauss, M. J. (2006). Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86(3), 572–588.

Varior, R. R., Haloi, M. & Wang, G. Gated siamese convolutional neural network architecture for human re-identification. *ECCV 2016*.

Viola, P. & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2), 137–154.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z. & Liu, W. (2018a). Cosface: Large margin cosine loss for deep FR. *CVPR*.

Wang, J., Zhou, F., Wen, S., Liu, X. & Lin, Y. (2017). Deep metric learning with angular loss. *CVPR*.

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on IP*, 13(4), 600–612.

Wang, Z., Tang, X., Luo, W. & Gao, S. (2018b). Face aging with identity-preserved conditional generative adversarial networks. *CVPR*.

Wanger, A., Wright, J., Ganesh, A., Zhou, Z. & Ma, Y. (2009). Towards a practical face recognition system: robust registration and illumination by sparse representation. *IEEE Trans. on PAMI*, 34(2), 597–604.

Wei, C.-P. & Wang, Y.-C. F. (2015). Undersampled face recognition via robust auxiliary dictionary learning. *IEEE Trans. on IP*, 24(6), 1722–1734.

Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A. K., Duncan, J. A., Allen, K. et al. (2017). IARPA Janus Benchmark-B face Dataset. *CVPR Workshop*.

Wong, Y., Chen, S., Mau, S., Sanderson, C. & Lovell, B. C. (2011). Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. *CVPR Workshop*.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S. & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Trans. on PAMI*, 31(2), 210–227.

Wu, X., He, R., Sun, Z. & Tan, T. (2018a). A light cnn for deep face representation with noisy labels. *IEEE Trans. on IFS*, 13(11), 2884–2896.

Wu, X., He, R., Sun, Z. & Tan, T. (2018b). A light cnn for deep face representation with noisy labels. *IEEE Trans. on IFS*, 13(11), 2884–2896.

Xie, W., Jia, X., Shen, L. & Yang, M. (2019). Sparse deep feature learning for facial expression recognition. *Pattern Recognition*, 96.

Xu, Y., Zhong, Z., Yang, J., You, J. & Zhang, D. (2017). A new discriminative sparse representation method for robust face recognition via l_2 regularization. *IEEE Trans. on Neural Networks and Learning Systems*, 28(10), 2233–2242.

Yang, M., Van Gool, L. & Zhang, L. (2013). Sparse variation dictionary learning for face recognition with a single training sample per person. *ICCV*.

Yang, M., Wang, X., Zeng, G. & Shen, L. (2017). Joint and collaborative representation with local adaptive convolution feature for face recognition with single sample per person. *Pattern Recognition*, 66, 117–128.

Yi, D., Lei, Z., Liao, S. & Li, S. Z. (2014). Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*.

Yin, B., Tran, L., Li, H., Shen, X. & Liu, X. (2019). Towards interpretable face recognition. *ICCV*.

Yin, X. & Liu, X. (2017). Multi-task convolutional neural network for pose-invariant FR. *IEEE Trans. IP*, 27(2), 964–975.

Yu, X., Fernando, B., Hartley, R. & Porikli, F. (2018). Super-resolving very low-resolution face images with supplementary attributes. *CVPR*.

Zhang, F., Zhang, T., Mao, Q. & Xu, C. (2018a). Joint pose and expression modeling for facial expression recognition. *CVPR*.

Zhang, H., Nasrabadi, N. M., Zhang, Y. & Huang, T. S. (2012). Joint dynamic sparse representation for multi-view face recognition. *Pattern Recognition*, 45(4), 1290–1298.

Zhang, K., Zhang, Z., Li, Z. & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10), 1499–1503.

Zhang, L. & Samaras, D. (2006). Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. *IEEE Trans. on PAMI*, 28(3), 351–363.

Zhang, L., Dou, P. & Kakadiaris, I. A. (2018b). Patch-based face recognition using a hierarchical multi-label matcher. *Image and Vision Computing*, 73, 28–39.

Zhao, J. & et al. (2017). Dual-agent GANs for photorealistic and identity preserving profile face synthesis. *NIPS*.

Zhao, J., Cheng, Y., Xu, Y., Xiong, L., Li, J., Zhao, F., Jayashree, K., Pranata, S., Shen, S., Xing, J. et al. (2018). Towards pose invariant face recognition in the wild. *CVPR*.

Zhu, P., Yang, M., Zhang, L. & Lee, I.-Y. (2014). Local generic representation for face recognition with single sample per person. *ACCV*.

Zhuang, L., Chan, T.-H., Yang, A. Y., Sastry, S. S. & Ma, Y. (2015). Sparse illumination learning and transfer for single-sample face recognition with image corruption and misalignment. *International Journal of Computer Vision*, 114, 272–287.