Face Recognition in Video Surveillance from a Single Reference Sample Through Domain Adaptation

by

SAMAN BASHBAGHI

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

MONTREAL, SEPTEMBER 27, 2017

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS:

Mr. Eric Granger, Thesis Supervisor
Department of Automated Manufacturing Engineering, École de technologie supérieure

Mr. Robert Sabourin, Co-supervisor
Department of Automated Manufacturing Engineering, École de technologie supérieure

Mr. Guillaume-Alexandre Bilodeau, Co-supervisor
Department of Computer and Software Engineering, Polytechnique Montréal

Mr. Stéphane Coulombe, President of the Board of Examiners
Department of Software and IT Engineering, École de technologie supérieure

Mr. Marco Pedersoli, Member of the jury
Department of Automated Manufacturing Engineering, École de technologie supérieure

Mr. Langis Gagnon, External Independent Examiner
Scientific Director, Centre de Recherche Informatique de Montréal

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON SEPTEMBER 12, 2017

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

First of all, I would like to express my special appreciation and gratitude to my adviser Prof. Eric Granger, you have been a great supervisor and friend for me. Thank you for your continuous supports, for your patience and motivation that helped me in all the time of my research and writing this thesis. I would also like to express my deepest gratitude to my co-advisors Prof. Robert Sabourin and Prof. Guillaume-Alexandre Bilodeau. They have thought me scientific and critical thinking skills. I am very grateful for helping me to grow as a research scientist and to keep my PhD thesis in the right track.

My gratitude also goes to the committee members: Profs. Stéphane Coulombe, Marco Pedersoli and Dr. Langis Gagnon for evaluating my thesis and providing constructive comments.

I would like to thank my friends and colleagues from the LIVIA, Miguel de la Torre, Christophe Pagano, Rafael Cruz, Moslem Ouled Sghaier, Luiz Gustavo, Idrissa Coulibaly, Marc-Andre Carbonneau and Francis Migneault who have been supporting me throughout my research technically and scientifically, and for helping me to quickly adapt to this new environment. Furthermore, I wish to thank my Iranian colleagues: Roghayeh Soleymani, Fania Mokhayeri, Amin Abbasnejad and Sajjad Abdoli who have been always supportive. I am thankful to you for sharing your great experiences and for all the discussions and fun we have had.

An special thanks goes to Javad Soltani Rad and his wife, I am very grateful for the many times you helped me get back on my feet to keep going during these years in this country. A big thank you to all of my friends and professors from Hamedan to Montreal, past and present. I would like to mention Alireza Kohansali, Mohammad Firouzfar, Amirhossein Habibian, Ali Khandani and Mostafa Parchami that inspired me throughout my life and academic studies.

Lastly, and more importantly, a warm and heartfelt thank you to my parents and my love who always encouraged me throughout my entire life. This thesis would not exist without their supports. To them, I dedicate this thesis.

RECONNAISSANCE DE VISAGES EN VIDÉOSURVEILLANCE À PARTIR D'UN ÉCHANTILLON DE RÉFÉRENCE UNIQUE À PAR L'ADAPTATION DE DOMAINE

Saman BASHBAGHI

RÉSUMÉ

Au cours des dernières décennies, la reconnaissance de visage (RV) a connu une attraction importante dans de nombreuses applications, telles que l'application de la loi, la médecine légale, le contrôle d'accès, la sécurité de l'information et la vidéosurveillance, en raison de sa nature cachée et non intrusive. Les systèmes RV spécialisés pour la vidéosurveillance cherchent à détecter avec précision la présence d'individus d'intérêt sur un réseau distribué de caméras vidéo dans des conditions de capture incontrôlées. Par conséquent, reconnaître les visages des individus ciblés dans un tel environnement est un problème complexe parce que l'apparence des visages varie en raison des changements de pose, d'échelle, d'illumination, d'occlusion, de flou, etc. La complexité de calcul est également une considération importante en raison du nombre croissant de caméras, le temps de calcul des algorithmes de détection de visage, de suivi d'objet et de classification à la fine pointe de la technologie.

Dans cette thèse, des systèmes adaptatifs sont proposés pour une RV fidèle à la vidéo, où un seul (ou très peu) échantillon de références de visage est disponible pour concevoir un modèle de visage de chaque individu d'intérêt. Cette situation correspond à un mode d'utilisation réel et courant dans les applications de surveillance à partir d'une liste de contrôle en raison du coût de la capture d'images de référence, de leur et faisabilité ardue et de la gestion complexe des modèles de visage en évolution dans le temps. De plus, le nombre limité de références faciales peut avoir une incidence défavorable sur la robustesse des modèles de visages dû aux faibles variations intra classes de ceux-ci, ce qui affecte par conséquent la performance des systèmes de RV sur vidéos. En outre, un défi spécifique pour la RV de type image-à-vidéo sont les différences perçues entre le domaine d'enregistrement, où les visages de référence de haute qualité sont acquises avec des conditions de capture contrôlées à partir de caméras fixes, et le domaine opérationnel, où les visages sont acquises à l'aide de caméras vidéo sujettes aux conditions de capture incontrôlées. Pour surmonter le défi introduit à partir d'un unique échantillon de visage par personne, 3 nouveaux systèmes sont proposés. Ceux-ci reposent sur des représentations multiples de visages et une adaptation de domaine pour assurer une RV fidèle à la vidéo. En particulier, cette thèse présentera 3 contributions qui seront sommairement présentées aux paragraphes qui suivront. Ces contributions seront décrites en plus grand détails aux chapitres correspondants.

Au chapitre 3, une approche multi-classificateurs est proposée pour une RV image-à-vidéo robuste basée sur des représentations de visage multiples et diverses de la référence image unique d'un même individu. Lors de l'enregistrement d'un individu d'intérêt dans le système, le visage de référence unique est toujours modélisé en utilisant un ensemble de classificateurs SVM basés sur des descripteurs extraits à partir de subdivisions différentes du visage de l'individu. Plusieurs techniques d'extraction de caractéristiques sont appliquées aux subdivisions isolées

dans l'image de référence pour générer un groupe de SVM diversifié qui fournit une robustesse contre les facteurs nuisibles courants (ex : variations d'éclairage et de pose). L'estimation des sous-ensembles de caractéristiques discriminantes, des paramètres des classificateurs, des seuils de décision et des fonctions de fusion d'ensemble est obtenue à l'aide d'une image de référence de haute qualité et d'un grand nombre de visages capturés dans une vidéo de qualité inférieure des individus non ciblés dans la scène. Lors de la mise en opération, le sous-ensemble de SVM le plus compétent est sélectionné dynamiquement en fonction des conditions de capture observées. Enfin, un algorithme de suivi de visage regroupe graduellement les visages capturés par personnes correspondantes apparaissant dans la scène, tandis que chaque ensemble spécifique à l'individu effectue une classification de visage. L'accumulation de scores correspondants par trajectoire de visage mène vers une RV spatio-temporelle robuste lorsque les scores d'ensemble cumulés dépassent un seuil de détection. Les résultats expérimentaux obtenus avec les bases de données Chokepoint et COX-S2V montrent une amélioration significative de la performance par rapport aux systèmes de référence, en particulier lorsque les ensembles spécifiques à chaque individu (1) sont conçus en utilisant des SVM exemplaires plutôt que des SVM à classe unique, et (2) exploitent la fusion au niveau des scores des SVM locaux (formés à l'aide des fonctionnalités extraites de chaque subdivision du visage), plutôt que d'utiliser soit la fusion au niveau de la décision ou au niveau des caractéristiques avec un SVM global (formés par une concaténation des descripteurs de caractéristiques extraits des subdivisions du visage).

Au chapitre 4, un système multi-classificateurs (SMC) efficace est proposé pour une RV fidèle à la vidéo en fonction des représentations multiples et de l'adaptation de domaine (AD). Un ensemble de classificateurs exemplaires SVM (e-SVM) par individu est ainsi conçu pour améliorer la robustesse face aux variations intra classes. Lors de l'enregistrement d'un individu cible dans le système, un ensemble de classificateurs est encore une fois utilisé pour modéliser chaque référence unique, où les descripteurs de visage multiples et les sous-espaces de caractéristiques sélectionnées aléatoirement permettent de générer un groupe diversifié de classificateurs pour chaque subdivision de visage. Pour adapter ces ensembles au domaine opérationnel, les e-SVM sont entraînés à l'aide des subdivisions de visage étiquetées et extraites de l'image de référence de l'individu d'intérêt contre celles extraites des images fixes de référence correspondant à plusieurs autres individus non ciblées, en plus des subdivisions de visages non étiquettées extraites à partir des trajectoires vidéos capturées par des caméras de surveillance. Pendant la phase opérationnelle, les classificateurs les plus compétents par visage de test donné sont sélectionnés dynamiquement et pondérés en fonction des critères internes prédéterminés avec l'espace de caractéristiques des e-SVM. Ce chapitre présentera également une étude de l'impact associée à l'utilisation de différents schémas d'entraînement pour l'AD, ainsi que l'utilisation d'un ensemble de validation de visages formé des images fixes d'individus non ciblées et des trajectoires vidéos d'individus inconnus dans le domaine opérationnel. Les résultats indiquent que le système proposé peut dépasser la précision des techniques utilisées dans la littérature, mais avec une complexité de calcul nettement inférieure.

Au chapitre 5, un réseau de neurones convolutif (RNC) profond est proposé pour faire face aux divergences observées entre les régions d'intérêt du visage isolées dans les images fixes et

celles sur vidéo pour une RV robuste. À cette fin, un auto-encodeur de visage RNC appelé FFA-CNN est entraîné à l'aide de régions d'intérêt fixes et sur vidéos à l'aide d'un apprentissage multi-tâches supervisé de bout en bout du réseau. Une nouvelle fonction de coût combinant une pondération des coûts liés aux pixels, à la symétrie et la conservation de l'identité est introduite pour optimiser les paramètres de ce réseau de neurones. Le système FFA-CNN proposé intègre à la fois un réseau de reconstruction et un réseau de classification entièrement connecté, où le premier reconstruit une région d'intérêt frontale bien éclairée avec une expression de visage neutre à partir d'une paire de régions d'intérêt vidéo non frontales de basse qualité, et où le second est utilisé pour comparer les représentations d'image fixe et sur vidéo pour fournir des scores de classification. Ainsi, l'intégration de la fonction de perte pondérée proposée avec une approche d'apprentissage supervisé de bout en bout permet de générer des visages frontaux de haute qualité et d'apprendre des représentations de caractéristiques de visage discriminatives similaires pour de mêmes identités données. Les résultats de simulation obtenus avec la compétition COX Face DB confirment l'efficacité de la technique FFA-CNN proposée pour obtenir des performances convaincantes par rapport aux systèmes RV de type RNC dans la littérature.

Mots clés: Vidéosurveillance, reconnaissance de visage, échantillon unique par personne, systèmes multi-classificateurs, méthodes par ensembles, SVMs exemplaires, méthodes de sous-espace aléatoires, adaptation de domaine, sélection dynamique de classificateur, architectures d'apprentissage profond, réseaux de neurones convolutifs

FACE RECOGNITION IN VIDEO SURVEILLANCE FROM A SINGLE REFERENCE SAMPLE THROUGH DOMAIN ADAPTATION

Saman BASHBAGHI

ABSTRACT

Face recognition (FR) has received significant attention during the past decades in many applications, such as law enforcement, forensics, access controls, information security and video surveillance (VS), due to its covert and non-intrusive nature. FR systems specialized for VS seek to accurately detect the presence of target individuals of interest over a distributed network of video cameras under uncontrolled capture conditions. Therefore, recognizing faces of target individuals in such environment is a challenging problem because the appearance of faces varies due to changes in pose, scale, illumination, occlusion, blur, etc. The computational complexity is also an important consideration because of the growing number of cameras, and the processing time of state-of-the-art face detection, tracking and matching algorithms.

In this thesis, adaptive systems are proposed for accurate still-to-video FR, where a single (or very few) reference still or a mug-shot is available to design a facial model for the target individual. This is a common situation in real-world watch-list screening applications due to the cost and feasibility of capturing reference stills, and managing facial models over time. The limited number of reference stills can adversely affect the robustness of facial models to intra-class variations, and therefore the performance of still-to-video FR systems. Moreover, a specific challenge in still-to-video FR is the shift between the enrollment domain, where high-quality reference faces are captured under controlled conditions from still cameras, and the operational domain, where faces are captured with video cameras under uncontrolled conditions. To overcome the challenges of such single sample per person (SSPP) problems, 3 new systems are proposed for accurate still-to-video FR that are based on multiple face representations and domain adaptation. In particular, this thesis presents 3 contributions. These contributions are described with more details in the following statements.

In Chapter 3, a multi-classifier framework is proposed for robust still-to-video FR based on multiple and diverse face representations of a single reference face still. During enrollment of a target individual, the single reference face still is modeled using an ensemble of SVM classifiers based on different patches and face descriptors. Multiple feature extraction techniques are applied to patches isolated in the reference still to generate a diverse SVM pool that provides robustness to common nuisance factors (e.g., variations in illumination and pose). The estimation of discriminant feature subsets, classifier parameters, decision thresholds, and ensemble fusion functions is achieved using the high-quality reference still and a large number of faces captured in lower quality video of non-target individuals in the scene. During operations, the most competent subset of SVMs are dynamically selected according to capture conditions. Finally, a head-face tracker gradually regroups faces captured from different people appearing in a scene, while each individual-specific ensemble performs face matching. The accumulation of matching scores per face track leads to a robust spatio-temporal FR when accumulated ensemble scores surpass a detection threshold. Experimental results obtained with the Chokepoint and COX-S2V datasets show a significant improvement in performance w.r.t. reference systems, especially when individual-specific ensembles (1) are designed using exemplar-SVMs rather than one-class SVMs, and (2) exploit score-level fusion of local SVMs (trained using features extracted from each patch), rather than using either decision-level or feature-level fusion with a global SVM (trained by concatenating features extracted from patches).

In Chapter 4, an efficient multi-classifier system (MCS) is proposed for accurate still-to-video FR based on multiple face representations and domain adaptation (DA). An individual-specific ensemble of exemplar-SVM (e-SVM) classifiers is thereby designed to improve robustness to intra-class variations. During enrollment of a target individual, an ensemble is used to model the single reference still, where multiple face descriptors and random feature subspaces allow to generate a diverse pool of patch-wise classifiers. To adapt these ensembles to the operational domains, e-SVMs are trained using labeled face patches extracted from the reference still versus patches extracted from cohort and other non-target stills mixed with unlabeled patches extracted from the corresponding face trajectories captured with surveillance cameras. During operations, the most competent classifiers per given probe face are dynamically selected and weighted based on the internal criteria determined in the feature space of e-SVMs. This chapter also investigates the impact of using different training schemes for DA, as well as, the validation set of non-target faces extracted from stills and video trajectories of unknown individuals in the operational domain. The results indicate that the proposed system can surpass state-of-the-art accuracy, yet with a significantly lower computational complexity.

In Chapter 5, a deep convolutional neural network (CNN) is proposed to cope with the discrepancies between facial regions of interest (ROIs) isolated in still and video faces for robust still-to-video FR. To that end, a face-flow autoencoder CNN called FFA-CNN is trained using both still and video ROIs in a supervised end-to-end multi-task learning. A novel loss function containing a weighted combination of pixel-wise, symmetry-wise and identity preserving losses is introduced to optimize the network parameters. The proposed FFA-CNN incorporates a reconstruction network and a fully-connected classification network, where the former reconstructs a well-illuminated frontal ROI with neutral expression from a pair of low-quality non-frontal video ROIs and the latter is utilized to compare the still and video representations to provide matching scores. Thus, integrating the proposed weighted loss function with a supervised end-to-end training approach leads to generate high-quality frontal faces and learn discriminative face representations similar for the same identities. Simulation results obtained over challenging COX Face DB confirm the effectiveness of the proposed FFA-CNN to achieve convincing performance compared to current state-of-the-art CNN-based FR systems.

Keywords: Video Surveillance, Watch-List Screening, Face Recognition, Single Sample Per Person, Multi-Classifier Systems, Ensemble Methods, Exemplar-SVMs, Random Subspace Methods, Domain Adaptation, Dynamic Classifier Selection, Deep Learning Architectures, Convolutional Neural Networks

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ALGORITHMS

LIST OF ABREVIATIONS

ASM	Active Shape Model
AUC	Area Under Curve
AUPR	Area Under Precision-Recall
CNN	Convolutional Neural Network
CCM-CNN	Cross-Correlation Matching
CD	Critical distance
CFR-CNN	Canonical Face Representation CNN
CE	Cross-entropy
DA	Domain Adaptation
DEC	Different Error Costs
DS	Dynamic Selection
DW	Dynamic Weighting
ED	Enrollment Domain
e-SVM	exemplar-SVM
Ee-SVM	Ensemble of e-SVMs
EM	Expectation Maximization
ESRC	Extended Sparse Representation Classification
ESRC-DA	Extended Sparse Representation Classification - Domain Adaptation
FE	Feature Extraction

FFA-CNN	Face-Flow Autoencoder CNN
FR	Face Recognition
FPR	False Positive Rate
FoV	Field of View
GAN	Generative Adversarial Model
HOG	Histogram of Oriented Gradients
KNN	K-Nearest Neighbors
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LERM	Learning Euclidean-to-Riemannian Metric
LLE	Local Linear Embedding
LPP	Locality Preserving Projection
LPQ	Local Phase Quantization
MCS	Multi-Classifier System
OC-SVM	One-Class SVM
OD	Operational Domain
PCA	Principal Component Analysis
pAUC	Partial Area Under Curve
PR	Precision-Recall
PSCL	Point-to-Set Correlation Learning

RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RSM	Random Subspace Method
SIFT	Scale Invariant Feature Transform
SRC	Sparse Representation Classification
SSPP	Single Sample Per Person
SURF	Speeded-Up Robust Features
SVDL	Sparse Variation Dictionary Learning
SVM	Support Vector Machine
TBE-CNN	Trunk-Branch Ensemble
TM	Template Matching
TPR	True Positive Rate
VS	Video Surveillance
VSKNN	Video Surveillance K-Nearest Neighbors

LISTE OF SYMBOLS AND UNITS OF MEASUREMENTS

$\mathbf{a}$	Positive ROI pattern
$\mathbf{a}_{i,p}$	ROI pattern of representation i at patch p
$\mathbf{a}_{i,k,r}$	ROI patch pattern of random subspace r using face descriptor k at patch i
α	Lagrangian constant
α_a, β_a	Regression parameters
b	Bias term
c_i	classifier i
$c_{i,k,r}$	classifier trained for random subspace r using face descriptor k at patch i
corr	Correlation layer
conv	Convolution layer
C_1, C_2	Regularization terms
C^+, C^-	Misclassification costs
C, E_j	Ensemble of classifiers for individual j
C_j^*	Set of the most competent classifiers for individual j
d_i	Decision for representation i.
d_j^*	Final decision for target person j
$\mathrm{dist}(\mathbf{t}, ST_j)$, d_{NT}	Distance of the probe from the target still and the nearest non-target ROIs
D	Validation set
D_{test}	Dataset of probe video ROIs

ξ_i	Slack variables
$E_s\left(I^S\right)$	Classification network function
$E_v\left(I^{p(t)}, I^{p(t+1)}\right)$	Reconstruction network function
f_k	Face descriptor k
f_χ	Learning function
f_{LPQ}	Label image with blur invariant LPQ
$\mathcal{F}$	Original feature space
F_j^l	Labeled face descriptors of individual j
F_v^u	Unlabeled face descriptors of unknown person v
$\mathbf{F_x}$	Vector of all pixel positions x at frequency $\mathbf{u}$
$F(\mathbf{u}, x)$	Short-term Fourier transform at frequency $\mathbf{u}$
F	The fully-connected classification network
G_j	Reference target still for individual j
H	Classifier hyperplane
iconv	Identity convolution
I^F	Frontal face
$I^{p(t)}$, $I^{p(t+1)}$	A pair of consecutive video faces
I^S	High-quality Frontal still ROI
I^T	High-quality still corresponding to I^P
K	Kernel function

l	Spacing stride pixels
L_{pixel}	Pixel-wise loss
$L_{symmetry}$	Symmetry loss
$L_{identity}$	Identity preserving loss
$\mathbf{m}_{i,p}^{j}$	Patch model of representation i at patch p for target individual j
m_cxn_c	Size of ROI patches
M_cxN_c	Resolution of still and video ROIs
M	Matching labels
N_{fd}	Number of face descriptors (feature extraction techniques)
N	Number of batches
N_a	Number of unknown non-target persons
N_c	Number of classifiers
N_d	Dimension of feature subspace
N_{ntd}	Number of calibration videos of unknown persons
N_{ntu}	Number of testing videos of unknown persons
N_{rs}	Number of random subspaces
N'_{rs}	Variable number of random subspaces
N_{sv}	Number of support vectors
N_v	Number of video trajectories
N_{wl}	Number of individuals of interest

N_x	Neighborhood of image $f(x)$ at pixel position x
p, n	Positive and negative classes
p_i	ROI patch i
pr_i	Decoder predictions i
ρ	Significance level
N_p	Number of patches
P_c	Compact pool of classifiers
P_g	Generic pool of classifiers
P_j	Pool of classifiers for individual j
P_j^l	Labeled ROI Patches of individual j
P_v^u	Unlabeled ROI Patches of unknown person v
$q_j(x)$	Scalar quantizer of jth component of the Fourier coefficients
r_s	Still representation
r_v	Video representation
R^n	Real-number space of n features
R_χ	Set of positive values of f_χ
$\mathcal{R}$, RS	Random subspace
RP	Pruned random subspace
R_j^l	Labeled random subspaces of individual j
R_v^u	Unlabeled random subspaces of unknown person v

$RA_{p,j}$	Ranking of patches for individual j
$RA_{s,j}$	Ranking of subspaces for individual j
s_r	Random feature subspace r
s_k^w	Weighted score of classifier c_k
$sub(I^T)^i$	Downsamples I^T
$\mathbf{sv}_{i,k,r}$	Support vector of random subspace r using face descriptor k at patch i
$S_{i,p}$	Matching score between $a_{i,p}$ and $m_{i,p}^j$
S_j^*	Final score for target person j
ST_j^l	Labeled still of individual j
SV_j	Set of support vectors for individual j
$\|SV\|$	Number of support vectors
$\mathbf{t}$	Probe ROI pattern
$\mathbf{T}_{i,p}$	ROI pattern of the target reference still
T_j^l	Labeled video trajectory of individual j
T_v^u	Unlabeled video trajectory of unknown person v
$\mathbf{u}$	Vector of frequencies
U	Number of negative samples
$\mathbf{V}_{i,p}^s$	ROI pattern of the support vector s
$\mathbf{w}$	Weight vector
w_k	Relative competence of the classifier c_k

$\mathbf{w}_{\mathbf{u}}$	Basis vector of the 2-D discrete Fourier transform at frequency $\mathbf{u}$
ω_i	weight for the ith loss function
W	Spatio-temporal window size
x_i	Training sample i
χ	Training dataset
y_i	Training label i
ϕ	Mapping function
$\gamma_{i,p}$	Predefined threshold for representation $a_{i,p}$
λ_1, λ_2, λ_3	Regularization parameters

INTRODUCTION

Biometric systems attempt to authenticate individuals for security purposes based on one or more unique biometric traits, such as face, iris, fingerprint, etc. Such systems enhance security over traditional authentication tools (e.g. identification cards and passwords), since these tools can be easily stolen or forgotten. Different applications of biometric can be broadly categorized into three main groups including: (1) verification, (2) identification and (3) screening. In the first group, identity claim of a subject needs to be confirmed by matching his/her biometric features against only its dedicated corresponding model stored in the system (one-to-one matching). Features of a subject are compared with a set of known individuals to retrieve his/her identity for identification (one-to-many matching). In the last group, unknown individuals in a relatively large population are compared to a limited number of target individuals (many-to-some matching). However, verification and identification can be also considered as close-set problems, while screening is an open-set problem.

Face recognition (FR) among different types of biometric applications has attracted many researchers during the past decades because, contrary to other biometrics like iris, finger- or palm-print, of its covert and non-intrusive nature that requires a low cooperation from individuals. FR systems are widely deployed in many decision support systems, such as law enforcement, forensics, access controls, information security and video surveillance (VS) (Jain *et al.*, 2004). FR systems allow to recognize individuals of interest based on their facial models, where the facial model is generated from facial regions of interest (ROIs) extracted from reference stills (videos) to perform for classification (De la Torre Gomerra *et al.*, 2015).

FR systems can be designed and assessed using three main scenarios w.r.t. the nature of training (reference) and testing (operational) data (Zhao *et al.*, 2003; Tan *et al.*, 2006): (1) still-to-still, (2) still-to-video and (3) video-to-video FR scenarios. In still-to-still FR scenario, ROIs extracted from still images of individuals of interest are employed as reference data to design a face model during enrollment, where other still images are used as operational data to perform recognition during operations. In still-to-video FR scenario, facial models are also designed using ROIs extracted from reference stills, while video streams are fed into the system to perform recognition. Finally, frames extracted from video streams are considered as both reference and operational data in video-to-video FR scenario (Pagano *et al.*, 2014).

FR systems for VS applications attempt to accurately recognize individuals of interest over a distributed network of cameras. In VS, capture conditions typically range from semi-controlled with one person in the scene (e.g. passport inspection lanes and portals at airports), to uncontrolled free-flow in cluttered scenes (e.g. airport baggage claim areas, and subway stations). Two common types of applications in VS are: (1) watch-list screening (that requires a system for still-to-video FR scenario), and (2) face re-identification or search and retrieval (that requires a system for video-to-video FR scenario) (Pagano *et al.*, 2014; De la Torre Gomerra *et al.*, 2015; Bashbaghi *et al.*, 2017a). In the former application, reference face images or stills of target individuals of interest are used to design facial models, while in the latter, facial models are designed using faces captured in reference videos. This thesis is focused on still-to-video FR, as required in watch-list screening under semi- and unconstrained VS environments.

During enrollment of target individuals, facial regions of interests (ROIs) are isolated in reference images that were captured under controlled condition, and used to design facial models. Then, during operation, the ROIs of faces captured in videos are matched against the facial models of each individual enrolled to the watch-list. In VS, a person in a scene may be tracked over several frames, and matching scores may be accumulated over a facial trajectory (a group of ROIs that correspond to the same high-quality track of an individual) for robust spatio-temporal FR. An alarm is triggered if accumulated matching scores linked to a watch-list individual surpasses an individual-specific threshold (Chellappa *et al.*, 2010).

Problem Statement

In still-to-video FR, still images of individuals are used to design facial models, in contrast with video-to-video FR, where facial models are designed from faces captured from video frames. The number of target references is limited in still-to-video FR applications, and the characteristics of the still camera(s) used for design significantly differ from the video cameras used during operations. Thus, still-to-video FR involves matching the face models obtained from reference stills against faces captured over a network of distributed surveillance cameras to accurately detect the presence of target persons.

Watch-list screening is a challenging application that relies on still-to-video FR, where face models must be designed prior to matching based on a single or very few reference ROIs isolated in a high-quality stills (e.g., mugshot or passport ID photo) (Bashbaghi *et al.*, 2014). In this thesis, a single high-quality reference still image captured with still camera under controlled conditions is matched against lower-quality faces captured with video cameras under uncontrolled conditions. There are significant differences between the appearances of still ROI(s) captured with still camera under controlled condition and ROIs captured with surveillance cameras, according to various changes in ambient lighting, pose, blur, and occlusion, and also camera inter-operability (Matta & Dugelay, 2009; Barr *et al.*, 2012). Thus, the facial models must be designed to be representative of the actual VS environments.

Although it is challenging to design robust facial models based on a single sample per person (SSPP), several approaches have addressed this problem, such as multiple face representations, synthetic generation of virtual faces, and using auxiliary data from other people to enlarge the training set (Bashbaghi *et al.*, 2014; Kan *et al.*, 2013; Kamgar-Parsi *et al.*, 2011; Yang *et al.*, 2013). These techniques seek to enhance the robustness of face models to intra-class variations. In multiple representations, different patches and face descriptors are employed (Bashbaghi *et al.*, 2014, 2017a), while 2D morphing or 3D reconstructions are used to synthesize artificial face images (Kamgar-Parsi *et al.*, 2011; Mokhayeri *et al.*, 2015). A generic auxiliary dataset containing faces of other persons can be exploited to perform domain adaptation (Ma *et al.*, 2015), and sparse representation classification through dictionary learning (Yang *et al.*, 2013). However, techniques based on synthetic face generation and auxiliary data are more complex and computationally costly for real-time watch-list screening applications, because of the prior knowledge required to locate the facial components reliably, and the large differences between quality of still and video ROIs, respectively.

Still-to-video FR systems proposed in the literature are typically modeled as individual-specific face detectors using one- or 2-class classifiers in order to enable the system to add or remove other individuals and adapt over time (Pagano *et al.*, 2014; Bashbaghi *et al.*, 2014). Modular systems designed using individual-specific ensembles have been successfully applied to the detection of target individuals in VS (Pagano *et al.*, 2014; De la Torre Gomerra *et al.*, 2015). Thus, ensemble-based methods have been shown as a reliable solution to deal with imbalanced data, where multiple face representations can be encoded into ensembles of exemplar-SVMs (e-SVMs) to improve the robustness of still-to-video FR (Bashbaghi *et al.*, 2015, 2017a). Multiple face representations of a single target ROI pattern has been shown to significantly improve the overall performance of basic template-based still-to-video FR system (Bashbaghi *et al.*, 2017a; Li *et al.*, 2013b). In particular, classifier ensembles can increase the accuracy of still-to-video FR by integrating diverse pools of classifiers. Furthermore, dynamic classifier selection methods allow to consider the most competent classifiers from the pool for a given face probe (Bashbaghi *et al.*, 2017b; Cruz *et al.*, 2015; Gao & Koller, 2011; Matikainen *et al.*, 2012). In this context, dynamic selection (DS) and weighting (DW) of the classifiers can be exploited, where the base classifiers trained using limited and imbalanced training data (Cavalin *et al.*, 2012, 2013). Spatio-temporal recognition considering high-quality tracks can be also exploited to enhance the robustness, where a tracker is employed to regroup ROIs of a same person into trajectories due to accumulation of ensemble predictions (Dewan *et al.*, 2016).

In addition, still-to-video FR systems can be viewed as a domain adaptation (DA) problem, where the data distributions between the enrollment and operational domains are considerably different (Patel *et al.*, 2015). Capturing faces in unconstrained environments and at several locations may translate to large discrepancies between the source and target distributions, due to camera field of view (FoV). Real-world scenarios for watch-list screening are most specially pertinent for unsupervised DA, because it is costly and requires human efforts to provide labels for faces in the target domain containing a large number of unknown individuals (Qiu *et al.*, 2014; Ma *et al.*, 2015). According to the information transferred between these domains, two unsupervised DA approaches are relevant for still-to-video FR: (1) instance-based and (2) feature representation-based approaches (Pan & Yang, 2010). The former methods attempt to exploit parts of the enrollment domain (ED) for learning in the operational domain (OD), while the latter methods exploit OD to find a desired common representation space that reduces the difference between domain spaces, and subsequently the classification error. Different unsupervised DA training schemes have been proposed in (Bashbaghi *et al.*, 2017c) to train an ensemble of e-SVMs for each individual of interest participated in the watch-list.

In general, methods proposed for still-to-video FR can be broadly categorized into two main streams: (1) conventional or shallow learning, and (2) deep learning or convolutional neural network (CNN-based) methods. The conventional methods rely on hand-crafted feature extraction techniques and a pre-trained classifier along with fusion, while CNN-based methods automatically learn features and classifiers cojointly using massive amounts of data. In spite of improvements achieved using the conventional methods, yet they are less robust to real-world still-to-video FR scenario. On the other hand, there exists no feature extraction technique that can overcome all the challenges encountered in VS individually (Bashbaghi *et al.*, 2017a; Huang *et al.*, 2015; Taigman *et al.*, 2014). Recently, several CNN-based solutions have been proposed to learn effective face representations directly from training data through deep architecture and nonlinear feature mappings (Sun *et al.*, 2013, 2014b; Chellappa *et al.*, 2016; Huang *et al.*, 2012; Schroff *et al.*, 2015). In such methods, different loss functions can be considered in the training process to enhance the inter-personal variations, and simultaneously reduce the intra-personal variations. They can learn non-linear and discriminative feature representations to cover the existing gaps compared to the human visual system (Taigman *et al.*, 2014), while they are computationally costly and typically require a large number of labeled data to train. To address the SSPP problem in FR, a triplet-based loss function have been introduced in (Parkhi *et al.*, 2015; Schroff *et al.*, 2015; Ding & Tao, 2017; Parchami *et al.*, 2017a,b) to discriminate between a pair of matching ROIs and a pair of non-matching ROIs. Ensemble of CNNs, such as trunk-branch ensemble CNN (TBE-CNN) (Ding & Tao, 2017) and HaarNet (Parchami *et al.*, 2017a) have been shown to extracts features from the global appearance of faces (holistic representation), as well as, to embed asymmetrical features (local facial feature-based representations) to handle partial occlusion. Moreover, supervised autoencoders have been proposed to enforce faces with variations to be mapped to the canonical face (a well-illuminated frontal face with neutral expression) of the person in the SSPP scenario to generate robust feature representations (Gao *et al.*, 2015; Parchami *et al.*, 2017c).

Objectives and Contributions

The objective of this thesis is to design adaptive still-to-video FR systems for robust watch-list screening that can accurately recognize target individuals of interest under unconstrained environments. According to the constraints of real-world watch-list screening applications, these systems need to be designed considering only a high-quality single reference still captured from the ED under controlled conditions, while they should be operated over low-quality videos captured from the OD under uncontrolled conditions. In addition, the facial models designed during enrollment of target individuals are required to compensate the lack of extra reference target samples (profile views of the target individual), to be representative of the operational scenes and also to be robust against various nuisance factors frequently observed in VS environments. Therefore, to adapt these facial models to the OD and to overcome the remarkable differences between the enrollment and operational domains, DA problem has to be addressed as well. Furthermore, these systems are expected to perform real-time under severe data imbalance situations during operations. The main contributions of this thesis rely on designing robust and adaptive still-to-video FR systems with SSPP through conventional and deep learning based methods. These contributions are organized into the following chapters:

- In Chapter 3, a new multi-classifier framework based on multiple and diverse face representations is presented, where an ensemble of SVM classifiers is exploited to model the single high-quality reference still of target individuals. A specialized 2-class classification technique is adopted that can be trained using only a single positive sample, where a large number of low-quality video faces of non-target individuals are utilized to estimate the classifier parameters, feature subsets, decision thresholds and fusion functions of ensembles (Bashbaghi *et al.*, 2017a).
- In Chapter 4, a new light-weight dynamic selection/weighting of classifiers is described in the context of multi-classifier system. Random subspace method and domain adaptation are exploited to generate multiple diverse representations and training classifiers, respectively. The impact of several combinations of data is assessed during training of the e-SVMs through unsupervised domain adaptation using non-target faces obtained from stills and video trajectories of unknown individuals in the enrollment and operational domains (Bashbaghi *et al.*, 2017b; Malisiewicz *et al.*, 2011a).
- In Chapter 5, a new deep CNN-based solution using autoencoders is developed to reconstruct frontal well-illuminated faces with neutral expression from low-quality blurry video faces, as well as, to generate domain-invariant feature representations. This network leverages a supervised end-to-end training approach using a novel weighted loss function, where still references and video faces from the both source and target domains are simultaneously

considered to address the domain adaptation and SSPP problems (Parchami *et al.*, 2017c; Dosovitskiy *et al.*, 2015).

Organization of Thesis

The block diagram of flow between chapters of this thesis is shown in Figure 0.1.

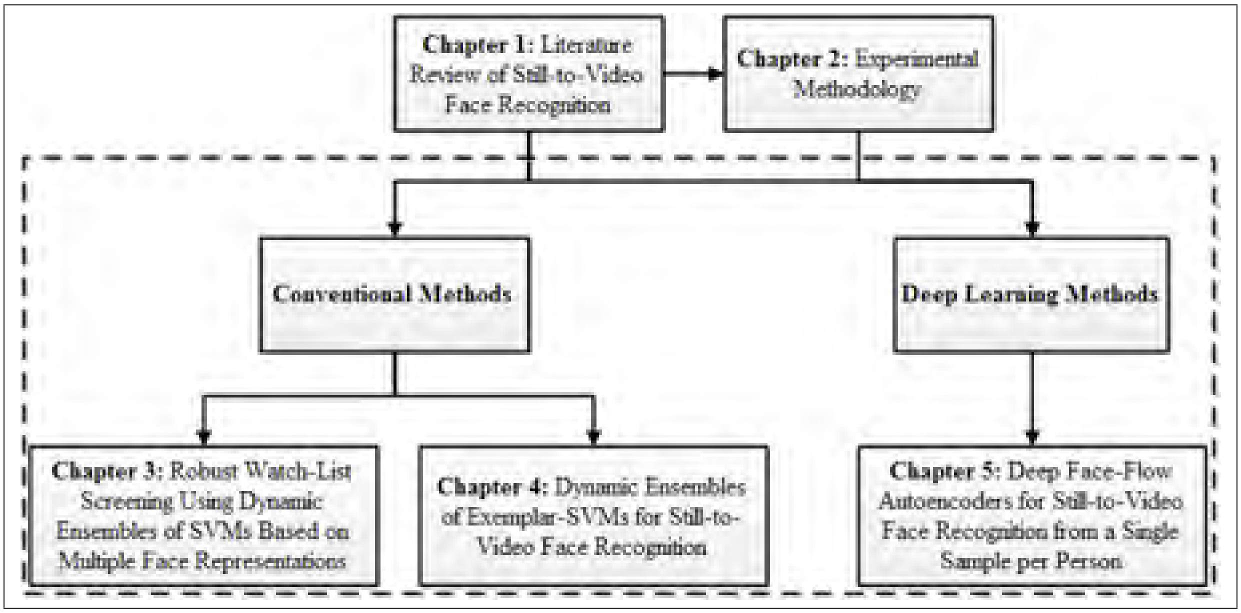

Chapter 1: Literature Review of Still-to-Video Face Recognition
Chapter 2: Experimental Methodology
Conventional Methods
Deep Learning Methods
Chapter 3: Robust Watch-List Screening Using Dynamic Ensembles of SVMs Based on Multiple Face Representations
Chapter 4: Dynamic Ensembles of Exemplar-SVMs for Still-to-Video Face Recognition
Chapter 5: Deep Face-Flow Autoencoders for Still-to-Video Face Recognition from a Single Sample per Person

Figure 0.1 The organization of this thesis.

The contents of this thesis are organized into four chapters. In Chapter 1, an overview of video-based FR literature in VS is presented focusing on still-to-video FR scenario. It starts with a generic still-to-video FR system, and followed by traditional and CNN-based state-of-the-art techniques have been proposed so far. The challenges of designing a robust still-to-video FR are discussed at the end of this chapter. In Chapter 2, the datasets and experimental methodology used to evaluate the proposed systems are described.

In Chapter 3, a multi-classifier framework is proposed that is robust for still-to-video FR when only one reference still is available during the design phase. The SSPP problem found in watch-list screening is addressed by exploiting multiple face representations, particularly through different patch configurations and several feature extraction techniques. Local Binary Pattern (LBP), Local Phase Quantization (LPQ), Histogram of Oriented Gradients (HOG), and Haar features are exploited to extract information from patches to provide robustness to local changes in illumination, blur, etc (Ahonen *et al.*, 2006, 2008; Bereta *et al.*, 2013; Deniz *et al.*, 2011). One-class support vector machine (OC-SVM) and 2-class exemplar-SVM are considered for the base classifiers, where the reference still and non-target videos are employed to train them, respectively. These specialized ensembles of SVMs model the variability in facial appearances by generating multiple and diverse face representations that are robust to various nuisance factors commonly found in VS environments, like variations in pose and illumination. Thus, SVM ensembles are trained using a single reference target ROI obtained from a high-quality generic still reference versus many non-target ROIs captured from low-quality videos. These non-target ROIs acquired from specific camera viewpoints, and of video cameras belonging to unknown people in the environment (background model) are used throughout the design process to estimate classifier parameters and ensemble fusion functions, to select discriminant feature subsets and decision thresholds, and to normalize the scores. To form discriminant ensembles, the benefits of selecting and combining patch- and descriptor-based classifiers with ensemble fusion at a feature-, score- and decision-level are also considered.

In Chapter 4, an efficient individual-specific ensemble of e-SVMS (Ee-SVMs) is proposed per target individual, where multiple face representations and domain adaptation are exploited to generate an E-eSVMs. Facial models are adapted to the OD by training the Ee-SVMs using a mixture of facial ROIs captured in the ED (the single labeled high-quality still of target and cohort captured under controlled conditions) and the OD (i.e., an abundance of unlabeled facial trajectories captured by surveillance cameras during a calibration process). Several training schemes are considered through DA for ensemble generation according to utilization of labeled ROIs in the ED and unlabeled ROIs in the OD. Semi-random feature subspaces corresponding to different face patches and descriptors are employed to generate a diverse pool of classifiers that provides robustness against different perturbations frequently observed in real-world surveillance environments. However, pruning of the less accurate classifiers is performed to store a compact pool of classifiers in order to alleviate computational complexity. During operations, a subset of the most competent classifiers is dynamically selected/weighted and combined into an ensemble for each probe using a novel distance-based criteria. Internal criteria are defined in the e-SVM feature space that rely on the distances between the input probe to the target still and non-target support vectors. In addition, persons appearing in a scene are tracked over multiple frames, where matching scores of each individual are integrated over a facial trajectory (i.e., group of ROIs linked to the high-quality track) for robust spatio-temporal FR. Experimental simulations with videos from the COX-S2V (Huang *et al.*, 2013a) and Chokepoint (Wong *et al.*, 2011) datasets indicate that the proposed system provides state-of-the-art accuracy, yet with a significantly lower computational complexity.

In Chapter 5, a supervised end-to-end autoencoder is proposed in this paper considering both still images and videos during the training of the network. In particular, a face-flow autoencoder CNN (FFA-CNN) is developed to deal with the SSPP problem in still-to-video FR, as well as, to restrain the differences between the enrollment and operational domains in the context of DA. The proposed FFA-CNN is trained using a novel weighted loss function to incorporate reconstruction and classification networks in order to recover high-quality frontal faces without blurriness from the low-quality video ROIs, and also to generate robust still and video representations similar for the same individuals through preserving identities to enhance matching capabilities. Therefore, the perturbation factors encountered in video surveillance environments and also the intra-class variations commonly exist in the SSPP scenario can be tackled using supervised end-to-end training. Simulation results obtained over challenging

COX Face DB (Huang *et al.*, 2015) confirm the effectiveness of the proposed FFA-CNN to achieve convincing performance compared to current state-of-the-art CNN-based FR systems.

Finally, general conclusions of this thesis and recommendations for future works are presented in the last Chapter.

CHAPTER 1

SYSTEMS FOR STILL-TO-VIDEO FACE RECOGNITION IN VIDEO SURVEILLANCE

This thesis presents a still-to-video FR system that can be employed as an intelligent decision support tool for VS. In surveillance applications, such as real-time screening of faces among a watch-list of target individual, the aim is to detect the presence of individuals of interest in unconstrained and changing environments. During enrollment of target individuals, facial models are designed using ROIs isolated in reference still images that were captured with a high-quality still camera 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 persons. A face tracker may be employed to track the subjects appeared in the capturing scene over several frames, and matching scores can be accumulated over a facial trajectory (a group of ROIs that correspond to the same high-quality track of an individual) for robust spatio-temporal FR (Chellappa *et al.*, 2010; De la Torre Gomerra *et al.*, 2015). This chapter presents a survey of state-of-the-art still-to-video FR systems and related techniques to address the SSPP and DA problems. In particular, methods for still-to-still FR and video-to-video FR are numerous but considered outside the scope of this thesis.

1.1 Generic Spatio-Temporal FR System

The generic system for spatio-temporal FR in VS is depicted in Figure 1.1.

As shown in Figure 1.1, each video camera captures the scene, where the segmentation and preprocessing module first detects faces and isolates the ROIs in video frames. Then, a face track is initiated for each new person appearing in the scene. Afterwards, feature extraction/selection module extracts an invariant and discriminative set of features. Once features are extracted, they are assembled into ROI patterns and processed by the classification module. Finally, classification allows to compare probe ROI patterns against facial models of individuals enrolled to the system to generate matching scores. The outputs of classification and tracking

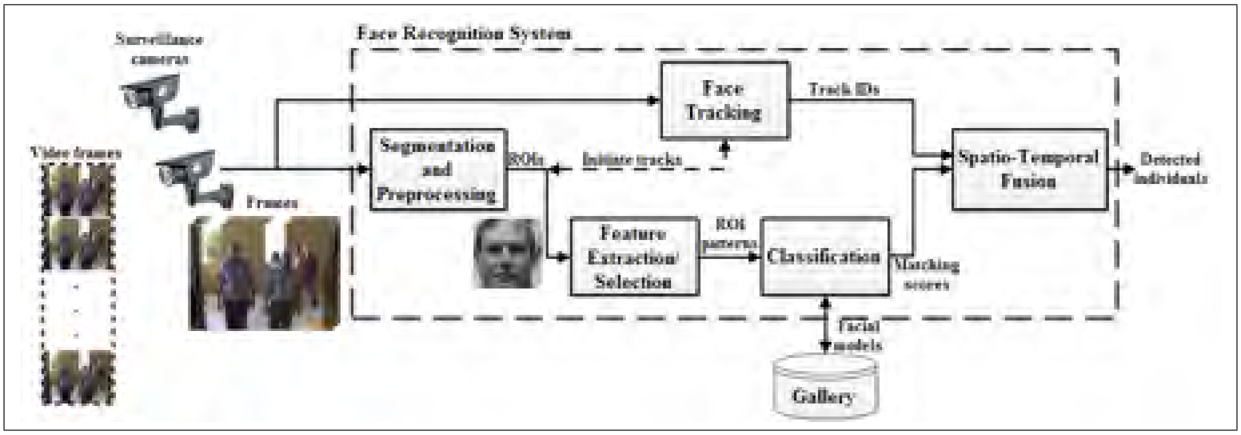


Figure 1.1 Generic system of spatio-temporal FR in video surveillance.

components are fused through spatio-temporal fusion module to achieve final detections (Chellappa *et al.*, 2010; Pagano *et al.*, 2012). This system is comprised of six main modules that are briefly described in the following items:

- Surveillance camera: Each surveillance camera in a distributed network of IP cameras captures the video streams of environment in its FoV that may contain one or more individuals appearing in the scene.
- Segmentation and preprocessing: The task of this module is detects faces from video frames and isolates the ROI(s). Typically, Viola-Jones (Viola & Jones, 2004) face detection algorithm is employed mostly due to its simplicity and speed. After obtaining the bounding box containing the position and pixels of face(s), histogram equalization and resizing of faces may be performed as the preprocessing step.
- Feature extraction/selection: Extracting robust features is an important step that converts ROI to a compact representation and it may improve the performance of recognition. Once the segmentation is carried out, some features are extracted from each ROI to generate the face models (for template matching). These features can be extracted from the entire face image (holistic) or local patches of it.

- Classification: After feature extraction, features are assembled into a feature vector (ROI pattern) and applied to the classification module. These features can be used in the simple template matcher or to train a statistical classifier for designing an appropriate facial model. Thus, recognition is typically performed using template matcher to measure the matching similarity between probe ROI and templates or using trained classifier to classify the input pattern to one of N predefined classes, each one belongs to enrolled watch-list individuals. In a still-to-video FR, one high quality still image (mug-shot) captured using a high resolution still camera is employed to design the facial model of each target individual of interest during enrollment, and then preserved in the gallery.

- Face tracker: This module regroups probe ROIs of a same individual captured over consecutive frames into facial trajectories by tracking the location of facial region of people appearing in the scene. It is beneficial for spatio-temporal recognition due to the accumulation the matching scores over time.

- Spatio-temporal fusion: Detecting the presence of target individuals can be achieved by combining matching scores of the classification and tracking modules. The spatio-temporal fusion can accumulate the output scores for each individual of interest over a window of fixed-size frames, and then compare the accumulated score over a trajectory with a predefined threshold (De la Torre Gomerra *et al.*, 2015).

1.1.1 State-of-the-Art Still-to-Video Face Recognition

There are many systems proposed in the literature for video-based FR, but very few are specialized for FR in VS (Barr *et al.*, 2012). Systems for FR in VS are typically modeled as a modular individual-specific detection problem, where each detector is implemented to accurately detect the individual of interest (Pagano *et al.*, 2012). Indeed, in these modular architectures adding and removing of individuals over time can be fulfilled easily, and also setting different decision thresholds, feature subsets and classifiers can be selected for a specific individual. Multi-classifier systems (MCS) are often used for FR in VS, where the number of non-target samples outnumbered target samples of individuals of interest (Bengio & Mariéthoz, 2007). An individual-specific approach based on one- or 2-class classifiers as a modular system with one detector for per individual has been proposed in (Jain & Ross, 2002). A TCM-kNN matcher was proposed in (Li & Wechsler, 2005) to design a multi-class classifier that employs a transductive inference to generate a class prediction for open-set problems in video-based FR, whereas a rejection option is defined for individuals have not enrolled to the system.

Ensembles of 2-class classifiers per target individuals were designed in (Pagano *et al.*, 2012) as an extension of modular approaches for each individual of interest in the watch-list for video-based person re-identification task. Thus, diversified pool of ARTMAP neural networks are co-jointly trained using dynamic particle swarm optimization based training strategy and then, some of them are selected and combined in the ROC space with Boolean combination. Another modular system was proposed based on SVM calssifiers in (Ekenel *et al.*, 2010) for real-time FR and door monitoring in the real-world surveillance settings. Furthermore, an adaptive ensemble-based system has been proposed to self-update the facial models, where the individual-specific ensemble is updated if its recognition over a trajectory is with high confidence (De la Torre Gomerra *et al.*, 2015).

A probabilistic tracking-and-recognition approach called sequential importance sampling (Zhou *et al.*, 2003) has been proposed for still-to-video FR by converting still-to-video into video-to-video using frames satisfying required scale and pose criteria during tracking. Similarly, a probabilistic mixture of Gaussians learning algorithm using expectation-maximization (EM) from sets of static images is presented for video-based FR system which is partially robust to occlusion, orientation, and expression changes (Zhang & Martínez, 2004). A matching-based algorithm employing several correlation filters is proposed for still-to-video FR from a gallery of a few still images in (Xie *et al.*, 2004), where it was assumed that the poses and viewpoints of the ROIs in video sequences are the same as corresponding training images. To match image sets in unconstrained environments, a regularized least square regression method has been proposed in (Wang *et al.*, 2015) based on heuristic assumptions (i.e. still faces and video frames of the same person are identical according to the identity space), as well as, synthesizing virtual face images. In addition, a point-to-set correlation learning approach has been proposed in (Huang *et al.*, 2015) for either still-to-video or video-to-still FR tasks, where Euclidean points are matched against Riemannian elements in order to learn maximum correlations between the heterogeneous data. Recently, a Grassmann manifold learning method has been proposed in (Zhu *et al.*, 2016) to address the still-to-video FR by generating multiple geodesic flows, to connect the subspaces constructed in between the still images and video clips.

Specialized feed-forward neural network using morphing to synthetically generate variations of a reference still is trained for each target individual for watch-list surveillance, where human perceptual capability is exploited to reject previously unseen faces (Kamgar-Parsi *et al.*, 2011). Recently, in (Huang *et al.*, 2013a) partial and local linear discriminant analysis has been proposed using samples containing a high-quality still and a set of low resolution video sequences of each individual for still-to-video FR as a baseline on the COX-S2V dataset. Similarly, coupling quality and geometric alignment with recognition (Huang *et al.*, 2013b) has been proposed, where the best qualified frames from video are selected to match against well-aligned high-quality face stills with the most similar quality. Low-rank regularized sparse representation is adopted in a unified framework to interact with quality alignment, geometric alignment, and face recognition. Since the characteristics of stills and videos are different, it could be an inefficient approach to build a common discriminant space. As a result, a weighted discriminant analysis method has been proposed in (Chen *et al.*, 2014) to learn a separate mapping for stills and videos by incorporating the intra-class compactness and inter-class separability as the learning objective.

Recently, sparse representation-based classification (SRC) methods have been shown to provide a high-level of performance in FR (Wright *et al.*, 2009). The conventional SRC method is not capable of operating with one reference still, yet an auxiliary training set has been exploited in extended SRC (ESRC) (Deng *et al.*, 2012) to enhance robustness to the intra-class variation. Similarly, an auxiliary training set has been exploited with the gallery set to develop a sparse variation dictionary learning (SVDL), where an adaptive projection is jointly learned to connect the generic set to the gallery set, and to construct a sparse dictionary with sufficient variations of representations (Yang *et al.*, 2013). In addition, an ESRC approach through domain adaptation (ESRC-DA) has been lately proposed in (Nourbakhsh *et al.*, 2016) for still-to-video FR incorporating matrix factorization and dictionary learning. Despite their capability to handle the SSPP problem, they are not fully-adapted for still-to-video FR systems. Indeed, they are relatively sensitive to variations in capture conditions (e.g., considerable changes in illumination, pose, and especially occlusion). In addition, samples in the generic training set are not necessarily similar to the samples in the gallery set due to the different cameras. Hence, the intra-class variation of training set may not translate to discriminative information regarding samples in the gallery set. They may also suffer from a high computational complexity, because of the sparse coding and the large and redundant dictionaries (Deng *et al.*, 2012; Yang *et al.*, 2013).

Video-based FR systems can make use of spatial information (e.g. face appearance) along with the location of persons and variations of faces over time to perform a robust spatio-temporal recognition. For instance, an adaptive appearance model tracking has been proposed for still-to-video FR (Dewan *et al.*, 2016) to learn track-face-model for each different individual appearing in the scene during operations. Sequential Karhunen-Loeve technique is employed within a particle filter-based tracker for online learning of track-face-models that are matched against the face models of individuals enrolled in the system. Moreover, A local facial feature based framework performing the matching of stills against video frames with different features (e.g., manifold to manifold distance, affine hull method, and multi-region histogram) has been proposed in (Shaokang *et al.*, 2011), where these features are extracted from a set of stills driven by utilizing spatial and temporal video information.

1.1.2 Challenges

In general, still-to-video FR as required in watch-list screening applications is a challenging problem. State-of-the-art FR systems perform poorly in the semi- or unconstrained environment, where the characteristics of still camera differ significantly from video surveillance cameras due to camera inter-operability (Best-Rowden *et al.*, 2013; Huang *et al.*, 2013a).

In addition, there are some nuisance factors that commonly observed in VS environments and can cause different variations in the appearance of the faces captured during operations (Matta, 2008). These factors include variations in pose, illumination, scale/distance, expression and imaging parameters as described below (Barr *et al.*, 2012):

- Pose: stationary cameras based on their FoV (viewpoints/angles) and also the locations of individuals may capture non-frontal faces with a variety of pose changes.
- Illumination: since each individual can pass through the cameras with different lighting conditions based on their position or ambient lighting and also their skin color, therefore, the lighting may vary and cause variety of face appearance at different time of the day.
- Scale: by moving individuals towards or away from the cameras, the face region will be larger or smaller in different video frames. So that, in the worst case, the face will become unrecognizable when it is very far or very close to the camera. However, the properties of camera such as depth of field of its lens may impact on the scale as well as the distance of the individual.
- Expression: facial expressions of individuals (e.g. happy, sad, angry, etc.) while passing through the camera may cause changing in the face appearance.
- Motion blur: blurriness can occur when the individual moves very fast or if the camera focus time takes too long (camera out of focus).
- Occlusion: when the other individuals or any objects in the capturing environment block parts of the face, the tasks of recognizing the face and distinguishing it from the background will be more difficult, especially in the crowded environment.

This problem becomes more difficult if only one single reference face is available for each person during the design. In this context, face models are not typically representative of faces captured in the operational environment. Nevertheless, it is important to extract multiple sources of information from just one available target sample. Estimating parameters of classifier with few design samples or validation set can lead to poor generalization and over-fitting. Furthermore, selecting representative non-target samples for each individual is needed to optimize performance due to overcome the issue of imbalanced data, defining thresholds, and also determining ensemble fusion functions.

1.2 Multiple Face Representations

Generating multiple face representations from the target reference still can improve robustness in watch-list screening applications. To compensate the unpleasant impacts of using only a single design reference, multiple face representations may be generated from the target reference still, using various feature extraction techniques and patch-based methods. Thus, to provide multiple and diverse representations w.r.t. the intrinsics of real-time still-to-video FR scenario, extracting different face descriptors and patches can be exploited. To that end, facial ROIs are first divided into several sub-regions (patches) with or without overlapping, then different feature extraction techniques (face descriptors) can be applied on each patch.

MCS specialized for spatio-temporal still-to-video FR contains individual-specific ensembles of classifiers generated for multiple face representations (see Figure 1.2). Facial ROIs in each frame are isolated using segmentation and preprocessing module. Meanwhile, the person tracker is initiated to regroup the facial ROIs captured for a same person into a trajectory. Then, multiple face representations are obtained by generating patterns that correspond to different patches and feature extractions to train a diverse pool of base classifiers. An individual-specific ensemble of classifiers is employed for multiple face representations. The fusion module combines the classification scores obtained using comparison of probe ROI pattern against facial models designed for each individual of interest.

1.2.1 Feature Extraction Techniques

Exploiting several discriminant face descriptors to generate multiple representations can be effective in still-to-video FR system. Each descriptor is specialized to address some nuisance

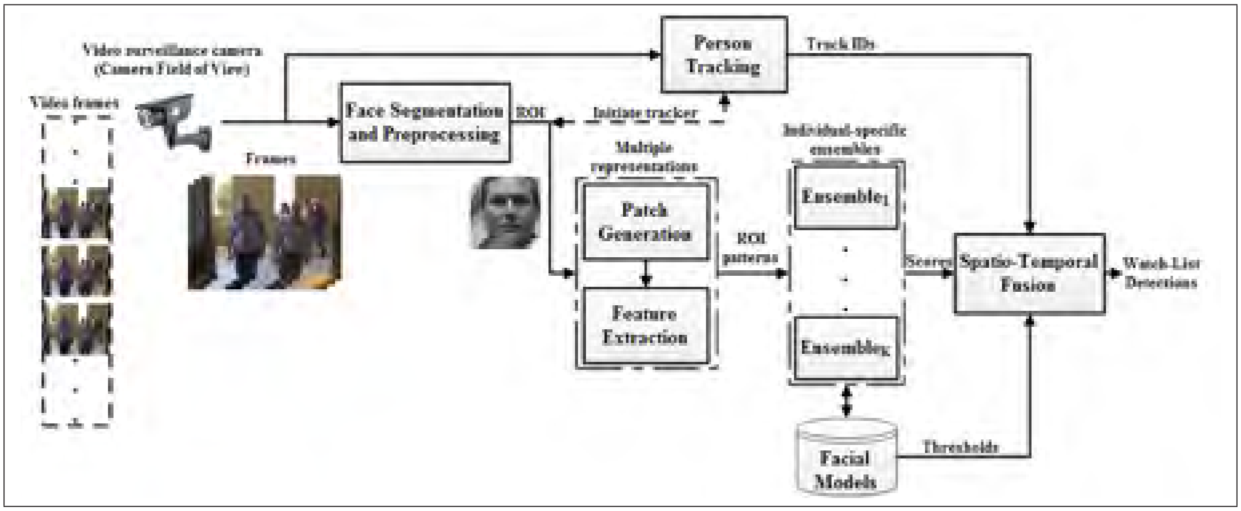


Figure 1.2 A multi-classifier system for still-to-video FR using multiple face representations.

factors (e.g., illumination, pose, blur, etc.) encountered in video surveillance. Hence, the choice of descriptors is based on the complementary information that they provide, where combining classifiers trained with different descriptors into an ensemble can achieve a high-level of robustness.

In FR literature, feature extraction techniques may be classified into holistic and local approaches based on locations and the ways they have been applied to face images (Abate *et al.*, 2007; Tan *et al.*, 2006).

- Holistic Approaches: These methods characterize the appearance of the entire face, and use the whole ROI to extract features. For instance, each ROI can be represented as a single high-dimensional ROI pattern by concatenating the grayscale (intensities) or color values of all pixels. In these appearance-based methods, all pixels of a face image may be involved in the extraction process. Holistic methods are generally divided into two main types as follows.

  a. Projection-based techniques: These methods typically transform the data from the original space to a new coordinate system in order to either reduce the dimensionality or classification process. Techniques such as PCA: principal component analysis (eigenface), LDA: linear discriminant analysis (fisherface), LPP: locality preserving projection (laplacianface), and LLE: local linear embedding (He *et al.*, 2003; Roweis & Saul, 2000; Zhang *et al.*, 1997) belong to this category. Due to the high-dimensional representation of face images, these techniques need sufficiently large training set to tackle the curse of dimensionality issue. Thus, they are not desired approaches to perform FR given a SSPP. However, they can be manipulated appropriately to provide either lower dimensional representation or feature selection.

b. Image processing techniques: In this category, image feature descriptors are exploited for providing face representation. These descriptors may scan either image regions and then extract features such as LBP or use image color histograms or mean/variance of grayscale values, and transformation such as Haar and Gabor (Ahonen *et al.*, 2006, 2008; Liu & Wechsler, 2002). Dense computation can be also applied to extract features from regions such as HOG (Deniz *et al.*, 2011).

- Local Approaches: These methods use local facial characteristics for generating face representation. Care should be taken when deciding how to incorporate global structural information into local face model. They are employed to characterize the information around a set of salient points, like eyes, nose, mouth, etc., or any local regions based on neighborhood or adjacency of pixels. They can be divided into two categories based on their definition of image locality.

a. Local facial feature-based techniques: These approaches first process the input image to locate and extract distinctive facial features such as the eyes, mouth, nose, etc., and then compute the geometric relationships among those facial points, thus reducing the facial image to a geometric feature vector. In other words, local facial features such as the eyes, nose, and mouth are taken into account along with their locations and local statistics (geometric and/or appearance). Therefore, these techniques extract structural information such as the width of the head, the distances between the eyes, etc. Thus, methods proposed based on extracting structural information aim to detect the eyes and mouth in real images, where various configurationally feature such as the distance between two eyes are manually derived from the single face image, such as Active Shape Model (ASM) (Zhao *et al.*, 2003).

b. Local appearance-based techniques: Local appearance-based methods extract information from defined local regions. Two steps are generally involved in these methods: (1) local region partitioning (to detect keypoints), and then (2) feature extraction from the neighborhood of those points. Local appearance-based face representation, like SIFT and SURF (Dreuw *et al.*, 2009) are generic local approaches and do not require determining of any salient local facial region manually.

1.2.2 Patch-Based Approaches

Patch-based methods allow to recognize faces in partially occluded unconstrained environments through local matching, where they may provide robustness to changes in pose and appearance (Liao *et al.*, 2013). Patch-based methods can be applied on the entire face image or local facial components (e.g., eye, nose, and mouth) of the face image (Lu *et al.*, 2013; Zou *et al.*, 2007). Patches can be defined uniformly using pyramid structures, saliency (detecting keypoints), or randomly. Local matching with patch-based methods potentially offer higher discrimination, allowing to recognize either partially occluded faces or arbitrary poses that appear frequently in unconstrained VS environments. Hence, patching makes use of local structural information to effectively deal with variations in uncontrolled surveillance conditions. Extracting features from local facial regions for local matching may lead to a robust and accurate FR systems.

1.2.3 Random Subspace Methods

RSMs randomly sample different feature subspaces from the original feature space of the input sample to create an ensemble of classifiers (Chawla & Bowyer, 2005). Let $\mathcal{F} = \{f_1, f_2, ..., f_d\}$ be the d-dimensional original feature space. To create a random subspace $\mathcal{R}$, s features are randomly sampled from $\mathcal{F}$. A feature vector belonging to the subspace $\mathcal{R}$ is denoted by $\mathbf{a} = [a_1, a_2, ..., a_s]$ and is used to train a classifier. This sampling process is repeated K times to create an ensemble of classifiers $C = \{c_1, \ldots, c_l, \ldots, c_K\}$, where using different subsets $\mathcal{R}$ encourage diversity among the classifiers c_l. The ensemble of classifiers C is therefore more suitable than a single classifier constructed with an instance from the complete feature space $\mathcal{F}$. Since RSM generates many redundant features, one of them may achieve higher accuracy compared to the original feature space. In the SSPP context, RSMs can provide different representations of the single training sample and inherit accuracy from classifier aggregation.

1.3 Domain Adaptation

When the training and test data are drawn from different distributions or feature spaces, classification performance can be typically magnified using knowledge transfer from the target domain with sufficient unlabeled data to the source domain with inadequate labeled data. To design a capable model for real-world applications with limited training labeled data, transfer learning would be a desirable strategy that avoids expensive efforts to collect and labeling the data. Thus, transfer learning aims to explore one or more source tasks to obtain knowledge in order to apply to a target task. In other words, different domains, tasks and distributions in training and testing are allowed through transfer learning. The importance of target task in transfer learning is higher than source task, due to capability of the model to operate on the target task (Pan & Yang, 2010).

Based on the task and data labels between the source and target domains, transfer leaning can be categorized into three settings comprised of (1) inductive and (2) transductive and (Pan & Yang, 2010). In inductive transfer learning, the target and source tasks are different, where some labeled data are available in the target domain. In transductive transfer learning, the source and target domains are different, while the task between the source and target is the same. Transductive transfer learning is associated with the situation that labeled data are available only in the source domain.

In transductive transfer learning as required in domain adaptation (DA), the learning function must be adapted using labeled data in the source domain, as well as, unlabeled data from the target domain. Transductive transfer learning setting can be carried out using approaches based on (1) instance-based transfer and feature representation-based transfer. The former approach is based on importance sampling and reweighting on the source domain data and then training models on the reweighted data (Quionero-Candela *et al.*, 2009), while the latter approach makes use of unlabeled data from the target domain to provide a feature representation across domains and learn a correspondence model (Blitzer *et al.*, 2006).

The key issue in DA is to learn a function f that can predict the class label of a novel input pattern regarding to changes in distribution of the source and target domain data. Domain adaptation problems can be defined as different approaches, such as semi-supervised, unsupervised, multi-source, and heterogeneous DA. In this regard, let X and Y denote the input (data) and the output (labels). Following Patel *et al.* (2015), let $S = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ denote the labeled data from the source domain, where $x^s \in X$ is an observation and $y^s \in Y$ is the corresponding class label. Labeled and unlabeled data from the target domain are denoted by $T_l = \left\{\left(x_i^{tl}, y_i^{tl}\right)\right\}_{i=1}^{N_{tl}}$ and $T_u = \{(x_i^{tu})\}_{i=1}^{N_{tu}}$, respectively. Thus, semi-supervised DA exploits the knowledge in S and T_l to learn the function f, while knowledge in S and T_u is used in unsupervised DA. In multi-source DA, function f may be learned from more than one domain in S along with any of T_l and T_u. Finally, heterogeneous DA can be considered when the dimensions of features in the source and target domains are not consistent.

Domain adaptation attempts to exploit a source domain with labeled data to learn a classification system for a target domain belonging to a different distribution. Two types of DA methods comprised of (1) semi-supervised and (2) unsupervised approaches have been studied based on availability of labeled data in the target domain. Unsupervised DA without any labeled data in the target domain is a more challenging problem than semi-supervised, while the latter approach leverages some labeled data in the target domain to conveniently provide associations between the both domains (Qiu *et al.*, 2014). Domain adaptation problems have been also addressed interchangeably under different concepts, such as covariate shift, class imbalance, and sample selection bias (Patel *et al.*, 2015).

It is therefore more sophisticated to tackle the problem of learning the similarity measure among data instances across domains in unsupervised DA approach. Structural correspondence learning (Blitzer *et al.*, 2006) is employed to model relations between the source and target data using pivot features that frequently appears in both domains in order to enforce correspondence among features from the two domains. Local geometry of data points in each domain is considered using a manifold-alignment based technique to compute the similarity between domains (Wang & Mahadevan, 2009). Maximum mean discrepancy is used in (Pan *et al.*, 2008) to measure the similarity between domains by learning a latent feature space.

Visual DA approaches can be categorized into the following strategies:

- In feature augmentation-based approaches, the key idea is to duplicate original features for each domain to make a domain-specific feature corresponds to both source and target domains. For instance, a common subspace (latent domain) has been introduced to compare the heterogeneous features from the source and target domains (Li *et al.*, 2014).

- In feature transformation-based approaches, the goal is to learn a transformation to adapt features across more general domains and use the learned similarity function to perform recognition (Baktashmotlagh *et al.*, 2013). This type of approaches is based on closeness between the target samples and the transformed source samples, where their computational complexity is high and depends on the number of training samples.

- In parameter adaptation methods, the general idea is to make use of kernel methods, where the objective function of a discriminative classifier is directly modified to transfer the adapted function for DA. For example, an adaptive SVM has been proposed in (Yang *et al.*, 2007) to adapt a source classifier which is trained on the source domain to a novel classifier for the target domain due to domain shift in videos.

- In dictionary-based approaches, the task is to learn an optimal dictionary and transfer it from one domain to the other domains, while maintaining low-dimensional or sparse representation characteristics of the dictionary on the new domain. A domain adaptive dictionary proposed in (Shekhar *et al.*, 2013) exploits a semi-supervised learning to build a single dictionary in order to represent both source and target domain data efficiently. Since the features are not correlated well in the original space, a common low-dimensional space has been considered to project the data from both domains and resolve the issue of correlation.
- In domain resampling methods, the key insight is that some samples in the source domain have much similarity than others to the instances of target domain. In (Gong *et al.*, 2013), an supervised DA method was developed based on picking out a subset of labeled data in the source domain that distribute the most similarity to the target domain in order to facilitate adaptation.
- In other methods, hierarchical DA approaches have been proposed to learn powerful non-linear representations of the data to incrementally capture information between the source and target domains using deep neural networks (Glorot *et al.*, 2011).

Domain adaptation methods can be typically applied on VS applications either in still-to-video FR or video-to-video FR scenarios. Capturing faces in unconstrained environments and different locations may cause several variations in the source and target distributions, due to different camera viewpoints, pose and illumination conditions, etc. However, real-world scenarios for face screening or re-identification are more pertinent to unsupervised DA, because it is costly and requires human efforts to provide labels for target faces. For instance, a dictionary-based DA approach has been proposed in (Qiu *et al.*, 2014) for video-to-video FR as required in re-identification of faces, where data in the source domain (early location) and target domain (final location) are drawn from different distributions. An unsupervised dictionary learning using intermediate domains along with domain-invariant sparse representation has been employed to link the source and target domains. Thus, intermediate subspaces have been synthesized to gradually reduce the reconstruction error of the target data. Similarly, finite or infinite number of intermediate subspaces is sampled to link the source and target domains to take into account the intrinsic domain shift (Gopalan *et al.*, 2011). Recently, a discriminative transfer learning approach has been proposed for the SSPP problem that relies on exploiting a generic training set (source domain) to learn a feature projection and then transfer into the single sample gallery set (target domain) through performing discriminant analysis (Hu, 2016). It attempts to minimize the differences between the source and target domains, and employs sparsity regularization to provide robustness against outliers and noise.

1.4 Ensemble-based Methods

One of the main approaches to address pattern recognition applications with limited and imbalanced training data is ensemble methods. The main idea of the ensemble is to generate several diverse classifiers over the original data, and to combine them through aggregating their predictions to outperform any single base classifier (Galar *et al.*, 2012; Skurichina & Duin, 2002). Thus, ensemble methods have been shown in many studies to improve the accuracy and robustness of a classification systems (Galar *et al.*, 2012; Granger *et al.*, 2012), where the accuracy and diversity of classifiers within ensembles are key issues in ensemble-based systems (Kuncheva & Whitaker, 2003; Zhu *et al.*, 2009). Accurate classifiers may provide a desirable performance, while simultaneously the classifiers need to be diverse from each other.

1.4.1 Generating and Combining Classifiers

To design an ensemble, a pool of diversified classifiers may be generated with training on different datasets or different parts of the input space. Every base classifier of the ensemble is a weak learner, where low changes in the data leads to large changes in the classification model (Galar *et al.*, 2012). To overcome the weakness of base classifiers, different techniques can be applied for the ensemble design with diversified classifiers, such sa Bagging, boosting and random subspace method (RSM). These are well-known re-sampling methods for ensemble design, where they first manipulate the training set, and training base classifiers on modified training sets, then, they combine classifier predictions into a final decision by adopting different ensemble fusion approaches (Skurichina & Duin, 2002; Kotsiantis, 2011).

Bagging introduced the concept of bootstrap aggregating, where different classifiers are trained using bootstrap replications over the original training data set. Bootstrapping is based on random sampling with replacement. Random sampling of instances (with original dimensions) from the training data set with replacement is formed to train each base classifier. Re-sampling produces different subsets that can guarantee the diversity of classifiers. In contrast to bagging, classifiers and training data are obtained in a deterministic way and sequentially in the boosting, not randomly and independently from the previous steps (Skurichina & Duin, 2002). As the second difference, boosting uses a function to produce a weight for voting, unless bagging applies equal weights for voting when combining classifiers (Kotsiantis, 2011). Thus, the ensemble of classifiers is induced by adaptively manipulating the distributions of training set based on the performance of the previously constructed classifiers. It uses the re-weighting of training data based on misclassified samples at each replication of boosting to generate the modified training set that leads to better performance. It is worth noting that bagging and boosting methods are not applicable to watch-list applications, because they require more than one target sample in the training set.

Random Subspace Methods allow for randomly sampling and selecting different feature subsets from the original feature space of input samples and then training several classifiers based on those subsets (Skurichina & Duin, 2002; Chawla & Bowyer, 2005). The dimension of subspaces sampled randomly is typically lower than the dimension of original feature space, where the training sample size increases relatively. Thus, it probably generates many redundant features that assist to obtain better classifiers. Each classifier constructed randomly by selecting subsets of a feature vector, where these subsets can prepare diversity by generating multiple representations. RSM can successfully apply to avoid over-fitting and it is more robust to noisy data. By sampling from the original feature space with lower dimensions, the number of training sample size can be increased, while the number of training samples is constant.

Therefore, when random subspaces have many redundant features, one of them may achieve higher accuracy compared to the original feature space (Skurichina & Duin, 2002).

In order to increase the classification accuracy, ensemble methods combine the outputs of several classifiers to provide the final output. Fusion techniques can be described in five levels (Connaughton *et al.*, 2013):

- Signal level: in this level, multiple input samples are combined in order to provide superior sample with higher information.
- Feature level: after feature extraction process, fusion is applied to combine all the extracted features into one feature vector.
- Score level: output scores that generated by different matchers/classifiers are fused at this level to produce the final result.
- Rank level: similar to score-level, this level of fusion combines match rankings instead of the output scores into a ranking scheme to define the best match.
- Decision level: in this level, Boolean responses of matchers for each sample are combined to obtain a final Boolean output by using Boolean operator or a voting method.

The most common fusion approach in FR systems is fusion at score-level (Connaughton *et al.*, 2013). In the score-level fusion, outputs of multiple matchers/classifiers are consolidated in order to compensate the performance of weak matchers/classifiers and also generate a new single scalar score. Fusion methods at this level can be organized into three different categories: density-based, classifier-based, and transformation-based schemes. Density-based fusion schemes are probabilistic approaches that approximate the density functions, such as Bayes decision rule or the minimum-error-rate classification rule. Classifier-based fusion schemes receive the scores of multiple matchers as input in order to train pattern classifier, such as neural networks for determining label of new samples. Finally, transformation-based fusion schemes combine the generated scores of multiple matchers/classifiers directly using simple fusion operators, such as the mean of scores or order statistics like minimum, maximum, and median of scores. Nevertheless, this approach would be meaningful only when the scores are comparable and have been normalized into a common domain. Typically, both density and classifier-based schemes require a large number of training samples (genuine and imposter scores) in order to accurately estimate the density function or computing the parameters of classifiers. Therefore, in case of limited number of training samples, transformation-based schemes can be the desired choice.

1.4.2 Classification Systems

Techniques for classifier design under class imbalance can be broadly categorized into: (1) algorithms that take into account the importance of positive samples (internal or algorithm-level), (2) techniques that apply preprocessing steps to re-balance the data distribution (external or data-level), and (3), cost-sensitive methods that combine both internal and external approaches to deal with different misclassification costs (Galar *et al.*, 2012). In addition, ensemble methods are often combined with one of the techniques above, specifically data-level and cost-sensitive ones (Li *et al.*, 2013b; Zhang & Wang, 2013).

For the design of classifier ensembles for watch-list screening under imbalanced data, specialized classification techniques are required. A simple approach for designing still-to-video FR system is using nearest neighbor classifier or template matching. In this case, the classification system performs hypothesis testing (or one-class classification) using a single reference face image per target individual. Template matching algorithms employ each facial model defined as a set of one or more templates stored in a gallery (Bereta *et al.*, 2013). It is also possible to consider a one-class classifier like Gaussian mixture modeling (Kemmler *et al.*, 2013) or one-class SVMs (Zong & Huang, 2011) to learn from an abundance of non-target class samples that are somehow similar to the single target class sample.

SVM is a widely used discriminative classifier that finds the optimal hyperplane to separate data patterns into two classes (Zeng & Gao, 2009). It requires a small number of training patterns to correctly model the boundary. Consider a training dataset $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$ in a 2-class classification problem, where $x_i \in R^n$ and $y_i \in \{-1, +1\}$ represent an n-dimensional data pattern and the classes of these data, respectively, for $i = 1, 2, \ldots, l$. These data patterns are typically mapped into a higher dimensional feature space using a mapping function ϕ to find the best separation of classes. Therefore, the soft-margin optimization problem is formulated as the following expression:

$$\min{}_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \tag{1.1}$$

$$\text{subject to } y_i\left(\mathbf{w}^T\phi\left(x_i\right) + b\right) \geq 1 - \xi_i,$$

$$\xi_i \geq 0$$

where slack variables ξ_i are introduced to account for misclassified examples. Thus, $\sum\limits_{i=1}^{l}\xi_i$ can be considered as a misclassification amount, **b** is the bias, and **w** is the weight vector. Constant C is a misclassification cost of a training example, where it controls the trade-off between maximizing the margin, as well as, minimizing the number of misclassifications.

Traditional SVM classifiers fail to classify imbalanced datasets properly, so that the estimated boundary is skewed to the majority class patterns (Batuwita & Palade, 2010). For classification of imbalanced datasets, the SVM objective function should be biased to push away the boundary from the majority class patterns in order to decrease the effect of class imbalance. The Different Error Costs (DEC) method (Veropoulos *et al.*, 1999) was proposed to modify the SVM objective function, where two misclassification cost values C^+ and C^- are assigned as follows:

$$\min{}_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^2 + C^+\sum_{[i|y_i=+1]}^{l}\xi_i + C^-\sum_{[i|y_i=-1]}^{l}\xi_i, \tag{1.2}$$

where C^{+} and C^{-} are the misclassification costs for the positive and negative classes, respectively. The optimal result is typically achieved when C^{+}/C^{-} equals the imbalance ratio (Zhang & Wang, 2013).

To overcome the class imbalanced issue and high misclassification rate, another SVM strategy named z-SVM is proposed to automatically orient the skewed decision boundary (Imam *et al.*, 2006). This method adjust the trained decision boundary toward the minority class regardless of learning parameters, contrary to existing SVM classifiers that exploits additional parameters empirically. To that end, a multiplicative weight z is assigned to each support vectors belonging to the minority class as follows:

$$f(x,z) = z \sum_{x_p \in SV;\left[p|y_p=+1\right]} \alpha_p y_p K(x,x_p) + \sum_{x_n \in SV;[n|y_n=+1]} \alpha_n y_n K(x,x_n) + b \tag{1.3}$$

where K is the kernel function, α is the Lagrangian constant, and SV is the set of support vectors. p and n indicates the positive and negative classes, respectively. This method is not convenient according to the assumptions of this chapter, because it requires more than one positive samples in the minority class.

In addition, one-class SVM (OC-SVM) classifier is designed to deal with data originating from only one class. It typically attempts to distinguish the samples of class of interest from all other outliers, where it defines a model using minimum volume contour (circle) to describe the target data (Krawczyk & Wozniak, 2014; Scholkopf & Smola, 2002). Basically, finding the optimal size of the volume is an indispensable issue due to the fact that a small volume may lead to an over-trained model, while a large volume size may accept outliers extensively.

Let $\chi = \{x_1, \ldots, x_m\}$ be the training dataset, where each x_j is a feature vector of a target sample. The goal of OC-SVM is to learn a function f_χ that assigns the data in χ to the set $R_\chi = \left\{f_\chi(x \geq 0)\right\}$, while minimizing the volume of R_χ. This issue is so called as estimation of minimal volume set, where membership of x to the set of R_χ determines whether its overall estimated volume is close to χ or not. A radial basis function (RBF) is used as a kernel function to estimate the minimal volume set. The OC-SVM constructs the hyperplane W to separate the training data mapped into the artificial feature space H from the hypersphere with radius equal to one $S_{(R=1)}$, as well as, to maximize the distance from it. Thus, the OC-SVM decision function $f_{\chi}(x)$ can be estimated as follows:

$$f_{\chi}(x) = \sum_{j}^{m} \alpha_j k\left(x_j, x\right) - \rho, \tag{1.4}$$

where $0 \leq \alpha_{\mathrm{j}} \leq \frac{1}{\mathrm{m}}$, $\sum_{\mathrm{j}} \alpha_{\mathrm{j}} = 1$, and the value of ρ is computed using $\mathrm{f}_{\chi}\left(\mathrm{x_j}\right) = 0$. Therefore, $\mathrm{x_j} \in \chi$ are located in the decision boundary when they satisfy both $\alpha_{\mathrm{j}} \neq 0$ and $\alpha_{\mathrm{j}} \neq \frac{1}{\mathrm{m}}$.

More recently, a method called exemplar-SVM (Malisiewicz *et al.*, 2011b) has been proposed to train a separate linear SVM classifier for every single positive example versus millions of negatives. The idea behind this method is to learn a separate 2-class classifier for each exemplar within a class of interest, unlike category-based classification. It is worth mentioning that this method has been mostly applied to ensemble learning in object detection and visual recognition tasks (Juneja *et al.*, 2013; Misra *et al.*, 2014). However, as described in Section 3.1, e-SVMs provide several advantages in the design of individual-specific ensembles for still-to-video FR. In particular, it can be trained with one target sample, and regardless of number of non-targets, as well as, it can rank the non-target support vectors w.r.t the target still. Furthermore, integrating multiple and diverse e-SVMs into an ensemble, with each e-SVM being specialized for a particular descriptor and facial zone provides a robust facial model.

Let **a** be the positive sample (target individual ROI pattern) and U be the number of non-target (negative) samples, respectively. The formulation of the e-SVM cost function is:

$$\min_{\mathbf{w},b} \left\{ \mathbf{w}^2 + C_1 \max(0, 1 - \left(\mathbf{w}^T a + b\right) + C_2 \sum_{x \in U} \max\left(0, 1 - \left(\mathbf{w}^T x + b\right)\right) \right\} \tag{1.5}$$

where C_1 and C_2 parameters control the weight of regularization terms, $\mathbf{w}$ is the weight vector, and b is the bias term. Since there is only one positive sample in the training set, its error is weighted much higher than the negative samples. The calibrated score of e-SVM for the given ROI pattern $\mathbf{a}$ and the learned regression parameters (α_a,β_a) is computed as follows:

$$f\left(x|\mathbf{w},\alpha_a,\beta_a\right) = \frac{1}{1+e^{-\alpha_a(\mathbf{w}_a^T-\beta_a)}}. \tag{1.6}$$

1.4.3 Dynamic Selection and Weighting of Classifiers

In general, the selection of diverse classifiers is a fundamental task in multi-classifier systems, where it can decrease the risk of classifier over-generalization by selecting a subset of accurate classifiers rather than all classifiers in the pool. The key idea of classifier selection is to select an ensemble of classifiers $C^* \in C$ that contains the most appropriate classifiers to classify a given input pattern a_i, where $C = \{c_1,\ldots,c_K\}$ is the pool of classifiers (K is the number of classifiers in the pool) and c_k is a base classifier in the pool C. This task can be basically categorized into static and dynamic classifier selection methods (Britto *et al.*, 2014). Methods that select the ensemble of classifiers statically are performed offline with a validation set, while dynamic selection (DS) methods exploit operating time contextual information (Misra *et al.*, 2014). The latter is preferred to select the most locally accurate set of classifiers based on context knowledge for each input pattern a_i. Dynamic weighting (DW) methods are similarly related to DS techniques, because they rely on the competence of classifiers (Galar *et al.*, 2015). A set of competent classifiers are dynamically selected from the ensemble to classify each input pattern in DS, while the scores of classifiers in the ensemble are weighted in DW.

A major issue to achieve a reliable DS/DW scheme are determining an accurate criterion to measure the level of competence among base classifiers c_k in the pool. The notion of competence as a selection approach indicates the capability of classifiers to best fit the given classifier selection process. In other words, it reveals a measure to select the best classifiers regarding to different classification tasks (Cruz *et al.*, 2015). It is worth noting that the Oracle as an abstract concept represents the ideal classifier selection strategy which always selects the classifier(s) that correctly classify the given input pattern in case of existing such classifier(s) in the pool, and rejects when there is no classifier that classify the input pattern correctly (Ko *et al.*, 2008). In other words, it provides the highest level of competence to the base classifier that predicts the correct label for a given input sample. Specifically, test sample a_i can be classified correctly by the pool, if at least one of the classifier in the pool can properly classify it.

In order to calculate the competence level of a base classifier, three different approaches were proposed in the literature so far containing: (1) the local neighborhood accuracy (over a region around the input test pattern a_i in the feature space), (2) decision templates or profiles (over a space declared by the output of base classifiers), and (3) extent of consensus.

a. Dynamic classifiers selection methods have been proposed in the literature are mostly based on the local accuracy concept (Britto *et al.*, 2014). Therefore, the accuracy of each classifier of the pool is estimated within the local region defined in a neighborhood of the pattern to be classified in the feature space (region of competence) (Didaci *et al.*, 2005). Typically, classifier accuracy is computed by employing k-NN to specify the neighborhood and then applying selection strategy to select the most accurate classifier. In the algorithm of overall local accuracy, the local accuracy of base classifiers is calculated as the correct classification percentage of the samples in the local neighborhood.

   Followed by (Didaci *et al.*, 2005), dynamic selection of an ensemble of classifiers called K-nearest-oracle (KNORA) is proposed in (Ko *et al.*, 2008) to optimize MCSs that select the most appropriate ensemble for each pattern instead of selecting the most accurate classifier. KNORA intuitively suggests different schemes including KNORA-ELIMINATE and KNORA-UNION to select the classifiers that correctly classify those similar neighbors in the validation set as the ensemble for classifying the given input pattern a_i. Despite the local accuracy concept, a greedy strategy can be also chosen to minimize the ensemble error in order to ensemble selection (Caruana *et al.*, 2006).

Since techniques using local accuracy to measure the competence highly rely on the performance of the methods employed to define the neighborhood, such as k-NN, they might fail in case of class segmentation dispersion and outliers. However, utilizing only local accuracy to measure the level of competence is not sufficient to yield the performance of Oracle. Furthermore, different distribution of validation and test set may slightly affect the dynamic selection system performance.

Diversity is considered as another important issue in the dynamic classifier selection task. Diversity among classifiers ensure that classifiers are suitable in different regions (e.g., feature space) to select based on the level of competence (Kuncheva & Whitaker, 2003). Hence, selecting the classifiers with higher level of competence surrounding local region of the given probe sample may lead to an accurate prediction.

b. In the decision templates techniques of defining competence, the similarity of the instance a_i is measured over the output space using the concept of decision output profile. The output decision profile of an input a_i can be denoted as $d_i = \left\{d_{i,1}, \ldots, d_{i,K}\right\}$, where each of $d_{i,k}$ is the output decision achieved using classifier c_k. Inspired by (Ko *et al.*, 2008), K-Nearest Output Profile (KNOP) fulfills dynamic ensemble selection (Cavalin *et al.*, 2012), where KNORA-Union is exploited to classify the input pattern. In this method, firstly the input test pattern is converted into output profile that comprised of the scores achieved by all the based classifiers and then the K-nearest output profiles in the database are selected using Euclidean distance. Finally, the group of classifiers that correctly classifies the samples in the validation set that their corresponding output profiles were already selected is chosen to classify the input pattern a_i.

As an advantage of decision templates techniques, it can be highlighted that they are not dependent on the quality of the region of competence in the feature space, while the decision space is considered to compute the similarity. Nevertheless, they only exploit global information of base classifiers instead of the local expertise of classifiers as a drawback.

c. Finally, a pool of ensemble of classifiers $C^{*'} = \left\{C_1^*, \ldots, C_{K'}^*\right\}$ (K' is the number of ensembles) is considered in techniques that rely on the extent of consensus rather than a pool of classifiers. Thus, the extent of consensus among base classifiers is used as the level of competence of the ensemble C_i^*. In such techniques, optimization algorithm like genetic algorithm or greedy strategy is employed to generate a population of ensemble of classifiers. For instance, criterion such as the margin between the first and the second voted class using the difference number of votes is proposed as an extent of consensus (Cavalin *et al.*, 2013). Another dynamic selection technique based on extent of consensus is Ambiguity-guided method, where the ambiguity among the classifiers of an ensemble is used as the level of competence criterion (Dos Santos *et al.*, 2008). The number of classifiers that disagrees with the overall ensemble decision is considered as the ambiguity. Thus, the lower the ambiguity value, the greater the level of competence of an ensemble.

Techniques based on extent of consensus are independent from the region of competence information, contrary to the local neighborhood accuracy techniques. However, since there are some ties among different members of the pool, an ensemble of classifiers with an acceptable consensus (level of confidence) may not be selected and the system may perform a random selection (Cavalin *et al.*, 2013). Moreover, the overall complexity of these techniques is higher than other aforementioned techniques, due to deal with a pool of ensemble of classifiers instead of a pool of classifiers.

Previous studies reveal that using only one criterion as a level of competence cannot be typically capable of selecting or weighting the classifiers dynamically and achieve a higher level of performance. However, multiple criteria can be considered to measure the competence of classifiers in order to appropriately select or weight them (Cruz *et al.*, 2015, 2017). For instance, a meta-learning framework was proposed in (Cruz *et al.*, 2015) to dynamic ensemble selection, where several distinct sets of meta-features are exploited to compute the level of competence among base classifiers. Each set of meta-feature takes different behavioral properties of a base classifier into account. Thus, the selection system can achieve suitable performance even if one criterion fails, due to other meta-features can be still considered throughout the selection scheme. Inspired from (Krawczyk *et al.*, 2014), an efficient selection method has been proposed in (Krawczyk & Cyganek, 2015) to automatically select locally specialized classifiers within ensembles of one-class classifiers due to overcome the complexity and time consuming drawbacks. In order to define several levels of competence, an optimal number of mutually complementary competence areas is determined. These competence areas are determined according to the clusters of one-class ensembles, where different groups of methods including the membership matrix, membership matrix and the dataset, and statistical indexes have been investigated to select number of clusters (number of competence areas).

1.5 Deep Learning Architectures

Deep convolutional neural networks (CNNs) have recently demonstrated a great achievement in many computer vision tasks, such as object detection, object recognition, etc. Such deep CNN models have shown to appropriately characterize different variations within a large amount of data and to learn a discriminative non-linear feature representation. Furthermore, they can be easily generalized to other vision tasks by adopting and fine-tuning pre-trained models through transfer learning (Schroff *et al.*, 2015; Chellappa *et al.*, 2016). Thus, they provide a successful tool for different applications of FR by learning effective feature representations directly from the face images (Chellappa *et al.*, 2016; Huang *et al.*, 2012; Schroff *et al.*, 2015). For example, DeepID, DeepID2, and DeepID2+ have been proposed in (Sun *et al.*, 2014a,b, 2015), respectively, to learn a set of discriminative high-level feature representations. In (Sun *et al.*, 2014b), an ensemble of CNN models was trained using the holistic face image along with several overlapping/non-overlapping face patches to handle the pose and partial occlusion variations. Fusion of these models is typically carried out by feature concatenation to construct over-complete and compact representations. Followed by (Sun *et al.*, 2014b), feature dimension of the last hidden layer was increased in (Sun *et al.*, 2014a, 2015), as well as, exploiting supervision to the convolutional layers in order to learn hierarchical and non-linear feature representations. These representations aim to enhance the inter-personal variations due to extraction of features from different identities separately, and simultaneously reduce the intra-personal variations.

In contrast to DeepID series, an accurate face alignment was incorporated in Microsoft DeepFace (Taigman *et al.*, 2014) to derive a robust face representation through a nine-layer deep CNN. In (Sun *et al.*, 2013), the high-level face similarity features were extracted jointly from a pair of faces instead of a single face through multiple deep CNNs for face verification applications. Similarly, for the SSPP problems, a triplet-based loss function has been lately exploited in (Schroff *et al.*, 2015; Parkhi *et al.*, 2015; Ding & Tao, 2017; Parchami *et al.*, 2017a,b) to learn robust face embeddings, where this type of loss seeks to discriminate between the positive pair of matching facial ROIs from the negative non-matching facial ROI. A robust facial representation learned through triplet-loss optimization has been proposed in (Parchami *et al.*, 2017b) using a compact and fast cross-correlation matching CNN (CCM-CNN). However, CNN models like the trunk-branch ensemble CNN (TBE-CNN) (Ding & Tao, 2017) and HaarNet (Parchami *et al.*, 2017a) can further improve robustness to variations in facial appearance by the cost of increasing computational complexity. In such models, the trunk network extracts features from the global appearance of faces (holistic representation), while the branch networks embed asymmetrical and complex facial traits. For instance, HaarNet employs three branch networks based on Haar-like features, while facial landmarks are considered in TBE-CNN. However, these specialized CNNs represent complex solutions that are not perfectly suitable for real-time FR applications (Canziani *et al.*, 2016).

Moreover, autoencoder neural networks can be typically employed to extract deterministic non-linear feature mappings robust to face images contaminated by different noises, such as illumination, expression and poses (Gao *et al.*, 2015; Parchami *et al.*, 2017c). An autoencoder network contains encoder and decoder modules, where the former module embed the input data to the hidden nodes, while the latter returns the hidden nodes to the original input data space with minimizing the reconstruction error(s) (Gao *et al.*, 2015). Several autoencoder networks inspired from (Vincent *et al.*, 2010) have been proposed to remove the aforementioned variances in face images (Gao *et al.*, 2015; Kan *et al.*, 2014; Le, 2013). These networks deal with faces containing different types of variations (e.g., illumination, pose, etc.) as noisy images. For instance, a facial component-based CNN has been learned in (Zhu *et al.*, 2014b) to transform faces with changes in pose and illumination to frontal view faces, where pose-invariant features of the last hidden layer are employed as face representations. Similarly, several deep architecture have been proposed using multi-task learning in order to rotate faces with arbitrary poses and illuminations to target-pose faces (Yim *et al.*, 2015; Zhu *et al.*, 2014a). In addition, a general deep architecture was introduced in (Ghodrati *et al.*, 2016) to encode a desired attribute and combine it with the input image to generate target images as similar as the input image with a visual attribute (a different illumination, facial appearance or new pose) without changing other aspects of a face. Moreover, a supervised deep architecture called FlowNet (Dosovitskiy *et al.*, 2015) has been proposed to solve the optical flow estimation, where feature vectors of a pair of images at different locations are correlated to accurately predict the other flows.

In the case of SSPP, a deep supervised auto-encoder has been proposed in (Gao *et al.*, 2015) to learn a robust face representation, where non-frontal faces with different nuisance factors are mapped toward the canonical face (frontal face with normal illumination and neutral expression) of the same person. In spite of their great success in FR with SSPP, they are not entirely desirable for still-to-video FR because of their computational complexity and also discrepancies in the domains of still and video images. To overcome these issues and tackle the constraints of DA, a simple canonical face representation through a supervised autoencoder (CFR-CNN) has been proposed in (Parchami *et al.*, 2017c) as the baseline still-to-video FR system that considers DA to reconstruct frontal faces from video ROIs. A fully-connected network was disjointly trained as a classifier to match the input probes. Subsequently, designing an accurate deep model requires to simultaneously consider both still images and videos during training and optimization of the network.

In addition, a supervised autoencoder has been proposed to compel faces with variations to be mapped to the canonical face (a well-illuminated frontal face with neutral expression) of the person in the SSPP scenario (Gao *et al.*, 2015). In contrast with standard autoencoders, this network was designed to extract similar features corresponding to the same persons to facilitate robust FR coupling with the conventional SRC in order to predict the labels of probe ROIs.

Lately, deep architectures using generative adversarial network (GAN) have been proposed for frontal view synthesis and also learn pose-invariant representations through an adversarial process (Huang *et al.*, 2017; Tran *et al.*, 2017). For example, global structure of the face and the transformation of local details are simultaneously handled by a two-pathway GAN, while prior knowledge from the distribution of frontal face is incorporated with a GAN to move a face with a large pose towards the frontal face. Since these approaches require landmark detection and also variations like blurriness and scale changes (distance of the person from surveillance cameras) are not considered, they are not fully adapted for video-based FR applications.

CHAPTER 2

EXPERIMENTAL METHODOLOGY

This chapter describes the experimental methodology used to evaluate the systems proposed in the thesis. It consists of video datasets, experimental protocols and performance measures adopted for the validation process.

2.1 Video Dataset

In real-world scenarios such as portals, the videos produced by surveillance cameras have some variations containing changes in illumination, pose, expression, and motion of individuals, scales and occlusion. Chokepoint (Wong *et al.*, 2011), COX-S2V (Huang *et al.*, 2013a) dataset are selected based on these characteristics (see Table 2.1). These video surveillance datasets can be employed to emulate real-world watch-list screening applications. The main characteristics of these datasets with respect to others (Beveridge *et al.*, 2013; Klare *et al.*, 2015) are that they contain a high-quality still face images captured under controlled condition (with the same still camera), and low-quality surveillance videos for each subject captured under uncontrolled conditions (with surveillance cameras).

Table 2.1 Characteristics of Chokepoint, COX-S2V datasets. Conditions include: indoor/outdoor (i/o); pose (p), illumination (l), expression (e), and scale (s); motion blur (b); occlusions (c); walking (w); random actions and/or motion (r); quality (v); and multiple people (m).

Characteristics	**Chokepoint**	**COX-S2V**
Number of persons	25 portal 1, 29 portal 2	1000
Resolution	800x600	1920x1080
Number of videos	54	4000
Frame Rate (fps)	30	25
Condition	i, p, l, e, s, b, c, w, v, m	i, p, l, e, s, b, w, v, b

Chokepoint dataset can be used as a benchmark for large-scale FR, especially in watch-list screening applications. An array of three cameras is placed above several portals to capture subjects walking through each portal in a natural way, used for simultaneously recording the entry or leaving of a person from different viewpoints (see Figure 2.1).

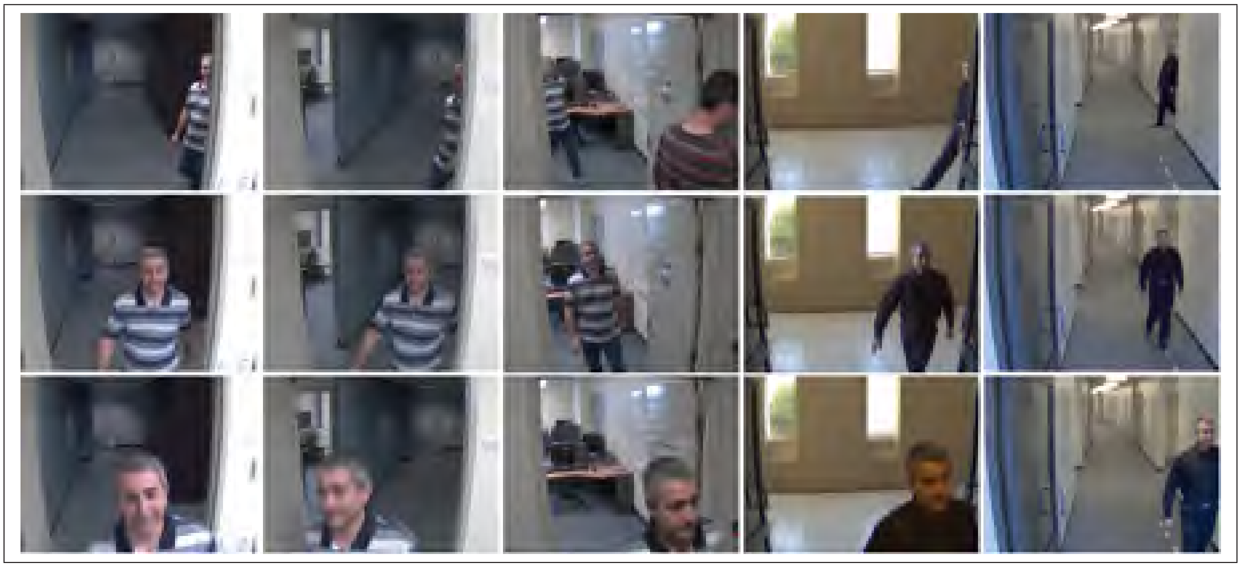


Figure 2.1 Example of video frames recorded by different cameras at several portals with various backgrounds in Chokepoint dataset.

While a person is walking through a portal, a sequence of face images can be captured. Random examples of neutral still ROIs of target individuals and ROIs captured from different trajectories are depicted in Figure 2.2. The variations between viewpoints allow for variations in walking directions, facilitating the capture of a near-frontal face by one of the cameras. In the database, each testing video sequence is named according to the recording conditions, for example (P1E_S1_C1) where P, S, and C stand for portal, sequence and camera, respectively. E and L indicate subjects entering or leaving the portal.

Another publicly available still-to-video dataset called COX-S2V dataset is also employed to fulfill more experiments on watch-list screening. This dataset contains high-quality controlled still faces of 1000 subjects along with uncontrolled low-quality video sequences, where video

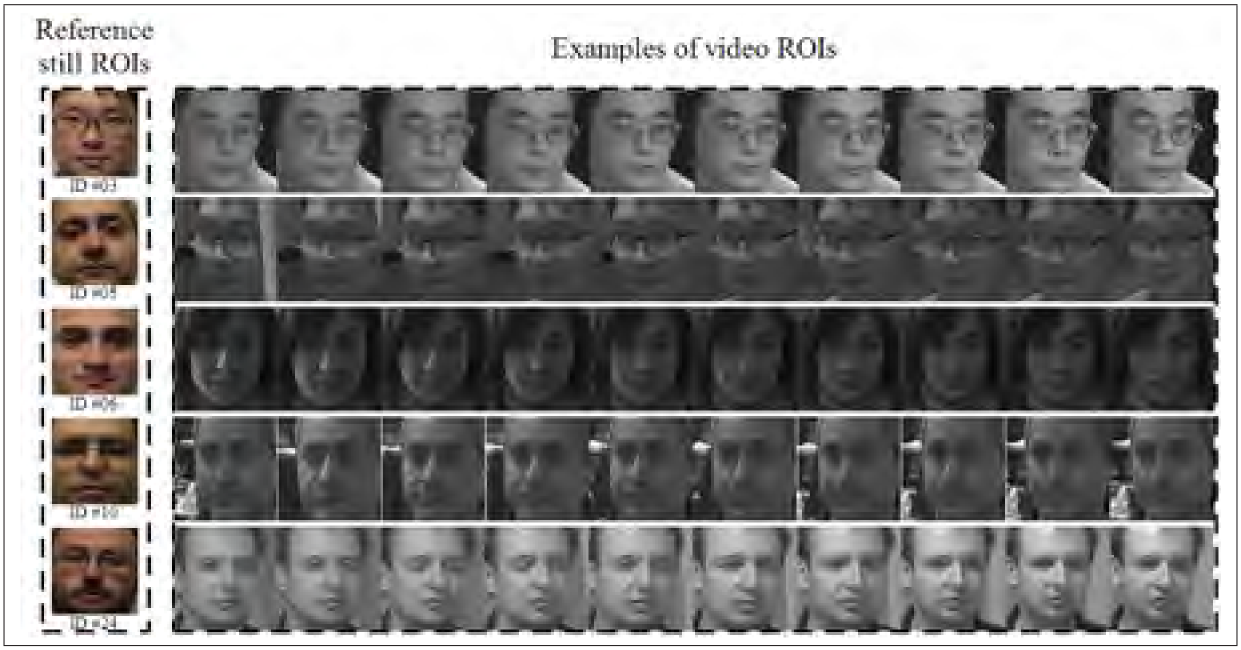


Figure 2.2 Example of ROIs captured from the 'neutral' mugshot of 5 target individuals of interest and random examples of ROIs captured from 5 different trajectories in Chokepoint videos.

clips are captured using two different off-the-shelf camcorders. In these videos, subjects walking naturally through a designed-S curve with changes in illumination, expression, scale, and viewpoint. Thus, four video clips with various resolutions are recorded per subject simulating VS scenario and located in the probe set.

An example of four low-resolution video sequences is shown in Figure 2.3. It is worth noting that this dataset is much more challenging than Chokepoint, because there are only 25 captures are available for each sequence during operation, as well as, the ROIs captured are blurry.

An example of one subject is demonstrated in Figure 2.4, showing the differences between ROIs captured in the enrollment and operational domains.

In addition, the IARPA Janus Benchmark A (IJB-A) (Klare *et al.*, 2015) is another dataset collected in the wild environment, where the still images and video frames were captured

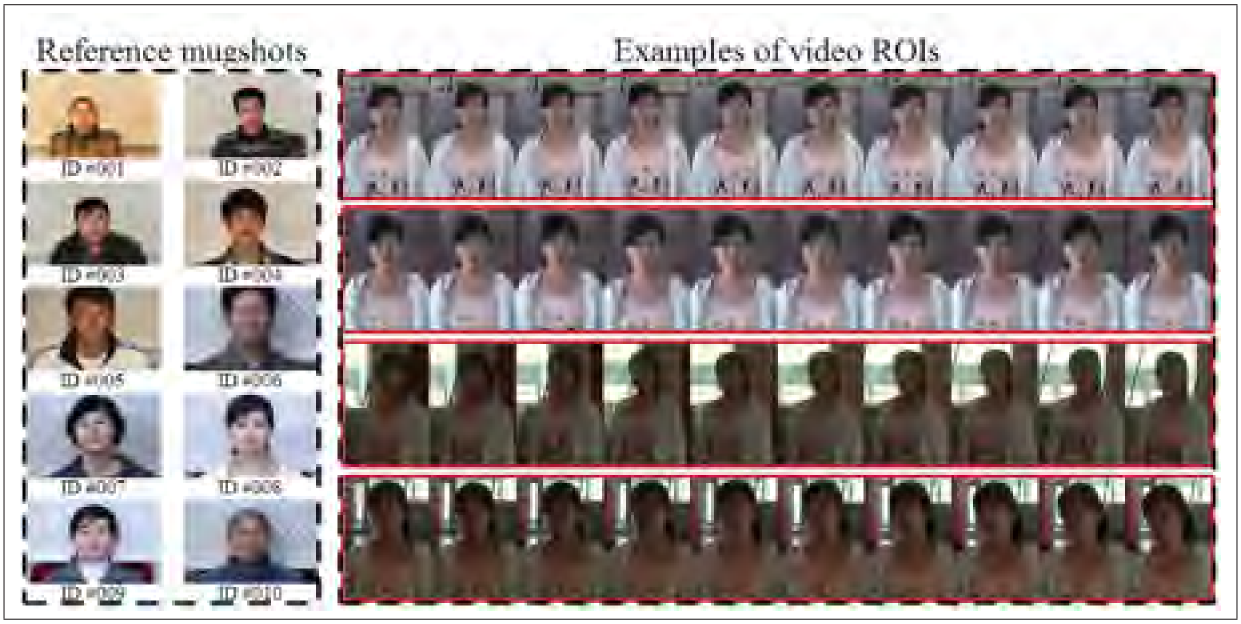


Figure 2.3 Example of individuals of interest enrolled in the watch-list and low-resolution video sequences captured with off-the-shelf camcorders under uncontrolled conditions in the COX-S2V dataset.

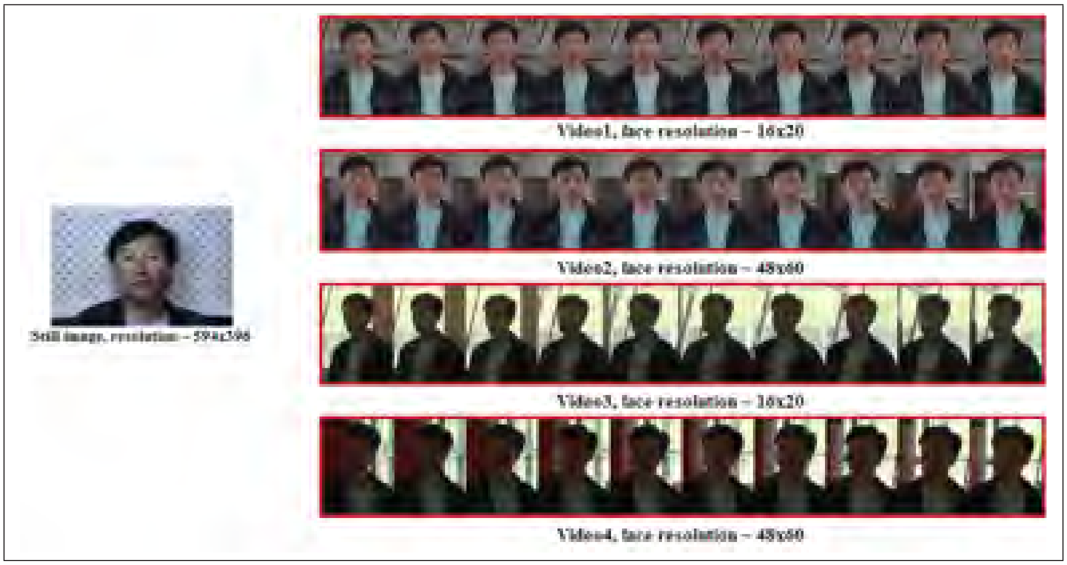

Video1, face resolution – 16x20
Video2, face resolution – 48x60
Still image, resolution – 594x396
Video3, face resolution – 16x20
Video4, face resolution – 48x60

Figure 2.4 An example of a still image belonging to one subject and corresponding four video sequences in the COX-S2V.

using different sensors within the wild setting under uncontrolled conditions. It is therefore not suitable for watch-list screening scenarios, because the still images are non-frontal and noisy.

2.2 Protocol for Validation

To validate the proposed systems in chapter 3 and 4, video sequences of Chokepoint data are chosen. In these experiments, 5 persons are selected randomly to be placed in the watch-list when only a single high-quality reference still and videos of 10 people that are assumed to come from non-target persons are used during enrollment. Thus, the rest of videos including 10 other non-target persons and 5 videos of persons who are already enrolled in the watch-list are used for testing. This process is repeated 5 times with random selection of targets and non-targets. Therefore, in each test iteration, target individuals (one at a time) and unknown individuals within the test videos pass through the portal, where the system seeks to detect the target person during operation and this process is repeated for other video sequences.

In the experiments using COX-S2V, 20 high-quality stills are randomly chosen from Persons-for-Publication folder (see Figure 2.3) to participate in the watch-list along with video clips recorded from all videos for design phase. Videos of 100 other persons are considered as non-target individuals for the design phase, as well as, videos of 100 other unknown persons as testing videos. Hence, one target individual at a time and non-target persons in the testing videos are participating in the operating phase. In order to accomplish statistically significant results, these experiments are iterated 5 times with 20 different individuals of interest.

2.3 Performance Metrics

The performance of still-to-video FR systems are typically assessed at the transaction-level to evaluate matching of Ee-SVMs for each ROI pattern (target versus non-target). Transaction-level analysis can be shown in the receiver operating characteristic (ROC) curves, in which true positive rates (TPRs) are plotted as a function of false positive rates (FPRs) over all threshold values. The proportion of target ROIs that correctly classified as individuals of interest over the total number of target ROIs in the sequence is considered as TPR. Meanwhile, FPR is the proportion of non-target ROIs incorrectly classified as individuals of interest over the total number of non-target ROIs. In a ROC space, a global scalar metric of the detection performance is the area under ROC curve (AUC), which can be interpreted as the probability of classification over the range of TPR and FPR. In other words, the AUC indicates correct ranking of positive-negative pairs in terms of class separation. For instance, AUC=100% shows an accurate discrimination among samples, where all positive are perfectly ranked higher than negatives.

In still-to-video FR system scenario, class priors of targets and non-targets may vary over time in each sequence. However, conventional ROC curves and AUC allow for evaluating the performance that is independent of mis-classification costs and class priors between classifiers. Thus, the precision-recall space can be employed in order to estimate the performance of the system at transaction-level, where it can characterize performance as the fraction of the correctly detected target ROIs against the total number of ROIs predicted belonging to an individual of interest. It is suitable to measure the system performance under highly imbalanced data situation during operations. Recall can be corresponded as TPR and precision (P) is computed as follows $\mathrm{P} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$.

In transaction-level analysis, performance of the watch-list screening system is provided using partial AUC (pAUC) and area under precision-recall (AUPR). Thus, pAUC(20%) is calculated using the AUC at $0 < FPR \leq 20\%$ in the ROC curve. The AUPR is desirable to illustrate the global accuracy of the system in the skewed imbalanced data circumstances. Experiments are iterated for each individual of interest in the watch-list for all video sequences, and then the average values are reported along with standard errors.

Moreover, the ground-truth track is employed to gradually group the captured ROIs over consecutive frames to create a trajectory due to trajectory-level analysis. To that end, captured ROIs of each individual in the operational scene are processed separately and the spatio-temporal fusion module accumulates ensemble scores over a window of fixed size to obtain the highest value inside the window in order to plot a ROC curve. Then the entire AUC is reported as a trajectory-level performance.

There exist several measures to estimate the ensemble diversity, that are computed based on classifier predictions (correct or incorrect for the class label) of the base classifiers. To assess the diversity of the proposed individual-specific ensembles of e-SVM classifiers, kappa (k) is calculated as a widely used diversity measure that is related to Kohavi-Wolpert variance and disagreement measures (Kuncheva & Whitaker, 2003). The value of k ranges from -1 to 1, where its lower values show greater diversity. The positive values indicate that the classifiers tend to classify the same object correctly, whereas the negative values correspond to negative correlation (Galar *et al.*, 2013).

CHAPTER 3

ROBUST WATCH-LIST SCREENING USING DYNAMIC ENSEMBLES OF SVMS BASED ON MULTIPLE FACE REPRESENTATIONS

The framework proposed in this chapter provides insights for the design of individual-specific ensembles that are robust in still-to-video FR when only one reference still is available to represent face models. Given the target and non-target data available for design, one-class Support Vector Machine (SVM) and the exemplar SVM (2-class) (Malisiewicz *et al.*, 2011b) are considered for the base classifiers. They follow a discriminant approach that is robust to limited reference data and class imbalance. These specialized ensembles of SVMs model the variability in facial appearances by generating multiple and diverse face representations that are robust to various nuisance factors commonly found in VS environments, like variations in pose and illumination.

During enrollment of a target individual, the corresponding facial model is encoded into an ensemble of specialized SVMs using a ROI extracted from a single high-quality reference still. A pool of diverse SVM classifiers is generated from multiple face representations of the reference ROI obtained by extracting face descriptors from patches. In particular, uniform non-overlapping patches are isolated in the reference ROI to improve robustness to occlusion (Liao *et al.*, 2013). Local Binary Pattern (LBP), Local Phase Quantization (LPQ), Histogram of Oriented Gradients (HOG), and Haar features are considered to extract information from patches to provide robustness to local changes in illumination, blur, etc (Ahonen *et al.*, 2006, 2008; Bereta *et al.*, 2013; Deniz *et al.*, 2011). During operations, ROIs of faces captured in videos are classified by each individual-specific ensemble of the system, and the ensemble scores are combined. Then, these scores are accumulated over trajectories of each person appearing in the scene for robust spatio-temporal recognition.

In particular, this chapter focuses on the analysis of different face patches and descriptors, one- and two-class SVM classifiers, and ensemble fusion strategies that are most suitable for the application constraints. Thus, SVM ensembles are trained using a single reference target ROI obtained from a high-quality generic still reference versus many non-target ROIs captured from low-quality videos. These non-target ROIs acquired from specific camera viewpoints, and of video cameras belonging to unknown people in the environment (background model) are used throughout the design process to estimate classifier parameters and ensemble fusion functions, to select discriminant feature subsets and decision thresholds, and to normalize the scores. To form discriminant ensembles, this chapter considers the benefits of selecting and combining patch- and descriptor-based classifiers with fusion at a feature-, score-, and decision-level, and by following a new dynamic selection strategy. To improve performance, specialized strategy allows to perform dynamic selection of ensembles based on patch ROIs with SVM properties, by measuring the distance between the target ROI patterns and support vectors.

The performance of the still-to-video FR systems designed according to the proposed framework are compared to reference systems (Bashbaghi *et al.*, 2014; Pagano *et al.*, 2014; Yang *et al.*, 2013) using videos from the publicly-available Chokepoint (Wong *et al.*, 2011) and COX-S2V (Huang *et al.*, 2013a) datasets. Accuracy and efficiency are measured at the transaction-level (matching of input probe ROI against reference ROI) and at the trajectory-level (the entire FR system over multiple frames). The contents of this chapter have been published in the journal of "Machine Vision and Applications" (Bashbaghi *et al.*, 2017a), the International Conference in Pattern Recognition (Bashbaghi *et al.*, 2014) and International Conference on Advanced Video and Signal Based Surveillance (Bashbaghi *et al.*, 2015).

3.1 Dynamic Ensembles of SVMs for Still-to-Video FR

A multi-classifier framework is proposed for robust still-to-video FR, where an ensemble of SVMs that encodes multiple discriminative face representations is assigned to each target individual (see Figure 3.1). Specifically, an individual-specific ensemble is designed using a diverse pool of specialized SVM classifiers to address the SSPP issue. This pool models the variability of faces by producing several face representations (various features extracted from patches) that are robust to common nuisance factors.

As illustrated in Figure 3.1, frames captured by a video camera may include several people. For each frame, preprocessing (e.g., grayscale conversion and histogram equalization) is first performed, and then segmentation is applied in order to isolate facial ROI(s). Then, the resulting ROIs are scaled into a predefined size and aligned based on the location of the eyes. Multiple ROI patterns are extracted from either the entire ROI, for i = 1, 2, ..., M number of feature extraction techniques, or from each patch p = 1, 2, ..., P. Each classifier (trained on the entire ROI or ROI patches) provides a matching score $S_{i,p}(\mathbf{a}_{i,p})$ between every ROI patch pattern $\mathbf{a}_{i,p}$ and the corresponding patch model $\mathbf{m}_{i,p^j}$ in the gallery index j=1, 2, ..., N indicates the number of individuals of interest enrolled to the system. Scores output from classifiers are fed into the fusion module after score normalization. A predefined threshold, $\gamma_{i,p}$ for each representation $\mathbf{a}_{i,p}$ is used to provide a decision d_i.

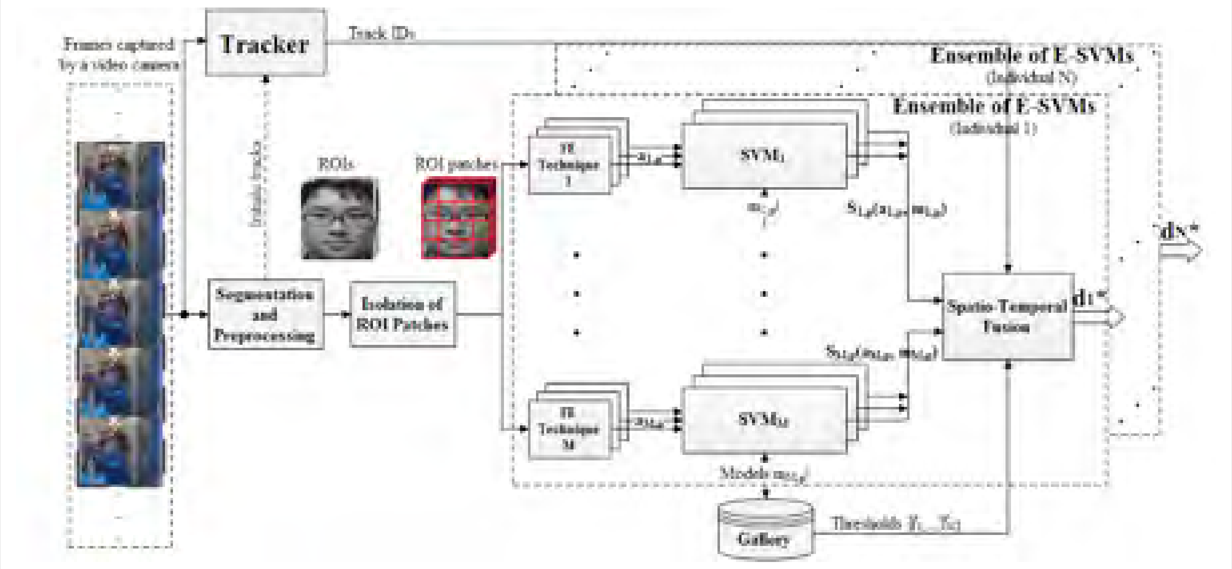


Figure 3.1 Block diagram of the proposed system for still-to-video FR.

In order to improve accuracy and robustness of recognition, each ensemble of SVMs is trained during enrollment with a single reference face still versus many of non-target faces captured from video cameras in the scene. Hence, a diverse pool of SVM classifiers is generated during design and then combined dynamically during operation to provide the ensemble score. Finally, ensemble scores are accumulated over trajectories defined using a face tracker to provide robust spatio-temporal recognition. The following subsections present details on the proposed system.

3.1.1 Enrollment Phase

During enrollment of a target individual, multiple face representations are generated from the ROI isolated in a single reference still, and in unlabeled facial trajectories of unknown non-target individuals. ROI patterns randomly extracted from videos of non-target individuals allow to select discriminant features, to train individual-specific ensemble of SVM classifiers, to define decision thresholds and to normalize the scores. Several SVMs are trained to estimate the face models based on each representation (features extracted on each uniform patches).

A single still reference is first converted to grayscale and then a facial ROI is isolated using a face detection method (Viola & Jones, 2004). Then, each ROI is scaled into a common predefined size, aligned, and then normalized for illumination invariance. Afterwards, different face descriptors are extracted from each patch in order to provide multiple face representations and generate a pool of diverse e-SVM and OC-SVM classifiers. For each representation, the ROI patch patterns of the target individual is combined with the corresponding ROI patterns of non-target individuals to train e-SVMs, while only ROI patterns of non-target individuals are employed to train OC-SVMs. For a system with P patches and M feature extraction techniques, enrollment involves generating a pool of MxP SVMs. Finally, decision thresholds computed from score distribution of non-targets (Bashbaghi *et al.*, 2014) and preserved in the gallery.

3.1.1.1 Extraction of Multiple Face Representations

Face descriptors (feature extraction techniques) and patch configurations employed in this chapter provide robustness to at least one of the nuisance factors that may occur in VS environments, such as changes in illumination, pose, and scale, providing multiple discriminant representations that are uncorrelated is important to design a robust watch-list screening system. Different feature extraction techniques from FR literature have been categorized in Table 3.1.

Table 3.1 The main nuisance factors for FR in VS and some feature extraction techniques that have been proposed to provide robust representations.

Illumination	LDA, Direct LDA, Kernel LDA, Kernel PCA, LBP, Gabor filters, RIU-LBP, LPP, Haar, SIFT, LQP, SURF, Daisy, HOG, LPQ
Expression	PCA, 2DPCA, Discriminant PCA, KPCA, LBP, LDA, Direct LDA, ICA, E(PC)2A, LPP, HOG, DCT, LPQ, Daisy
Pose	Direct LDA, Haar, HOG, LPQ
Rotation	LBP, Gabor filters, SIFT, HOG, SURF
Occlusion	HOG, Haar/SURF (partial occlusion)
Scale	SIFT, SURF, Daisy, HOG
Motion Blur	LPQ
Aging(Time)	LBP, 2DPCA, ICA, DCT

LBP (Ahonen *et al.*, 2006) and LPQ (Ahonen *et al.*, 2008) are popular face descriptors that extract texture features of faces in different way. LPQ is more robust to motion blur because it relies on the frequency domain (rather than spatial domain) through the Fourier transform. LBP preserves the edge information, which remains almost the same regardless of illumination change. HOG and Haar features are selected to extract the information more related to shape. HOG (Deniz *et al.*, 2011) is able to provide a high level of discrimination on a SSPP because it extracts edges in images with different angles and orientations. Furthermore, HOG is robust to small rotation and translation. Wavelet transforms have shown convincing results in the area of FR (Amira & Farrell, 2005). In particular, Haar transform performs well with respect to pose changes and partial occlusion.

Finally, using multiple face representations generated through different face descriptors extracted from every face patch can increase the diversity among classifiers, robustness to variations, and tolerance to some occlusions. For each patch, a classifier is trained with the reference still patch versus the corresponding patches of non-target faces captured among universal background model. Features extracted from non-overlapping uniform patches from each ROI are used to train classifiers.

3.1.1.2 Generation of Diverse SVM Classifiers

Given only one target reference still ROI captured under controlled condition (from another scene and camera), and an abundance of non-target ROIs captured from videos, training classification system to address the variabilities in VS environment is challenging. Thus, a framework with an ensemble per person is considered. They have been shown to provide robust and accurate performance when training data is limited (De la Torre Gomerra *et al.*, 2015). It is however challenging to train or generate a diverse pool of classifiers per target individual from the original data (Li *et al.*, 2013b).

In the SSPP problems, OC-SVMs can be trained considering only the non-target samples obtained from unknown individuals (Figure 3.2 ((a))), while e-SVMs can be trained using a single target sample (still ROI pattern) along with many non-target samples (video ROI patterns) for each individual of interest as illustrated in Figure 3.2 ((b)). Thus, training can be performed by considering non-target ROIs as negative samples obtained from background model. Subsequently, the information of non-target individuals from the field of view may be exploited during training to enhance the capability to generalize during operation.

The diversity of SVMs in a pool is produced using multiple representations. It should be noted that the input features must be normalized between 0 and 1 through min-max normalization performed based on non-target face samples. Normalization of output scores for fusion is taken into account using min-max normalization as well.

E-SVMs possess some potential benefits in designing individual-specific classifier systems with multiple face representations from only one positive versus several negative samples. The large number of non-target samples appears to constrain the SSPP problem. Since this classifier finds support vectors that are highly similar to each individual when training e-SVMs, the amount of negative sample cannot affect the accuracy of the decision boundary (Malisiewicz

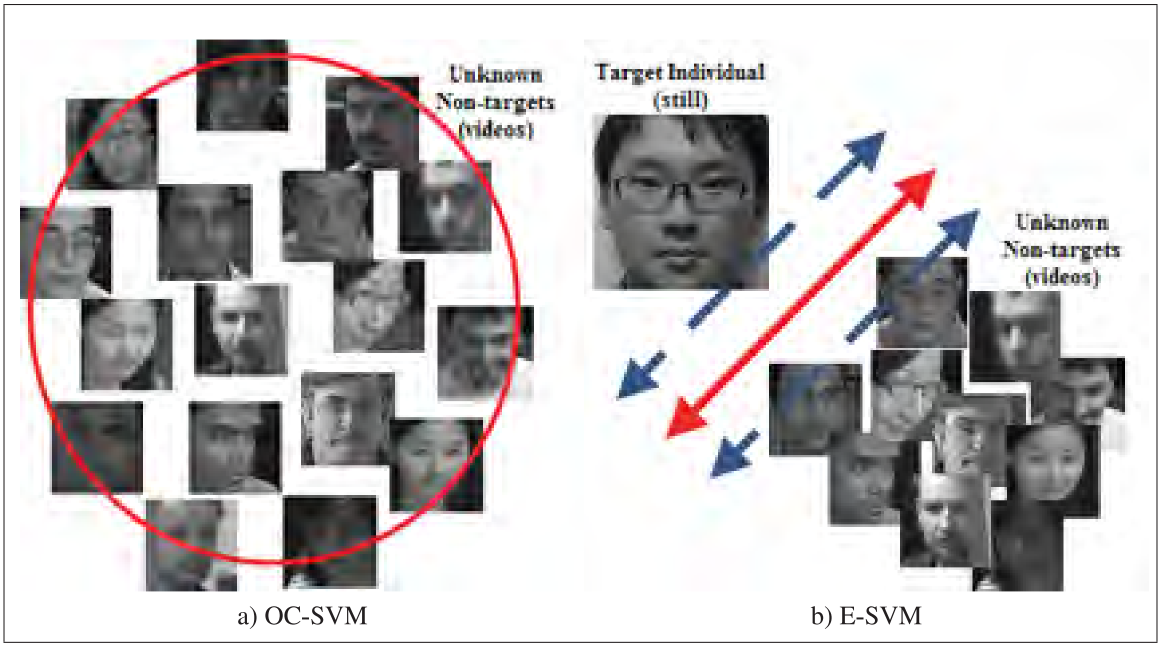


Figure 3.2 Illustration of training OC-SVM and e-SVM for each individual of interest enrolled in the watch-list with a single ROI block.

et al., 2011b). Hence, it can be applied suitably even for large databases containing few exemplars in the training set, e.g., as acquired in watch-list screening.

Since each e-SVM is highly specialized to the target individual, the largest margin (decision boundary) will be obtained by training under imbalanced data exploiting different regularization parameters, which provides more freedom in defining the decision boundary. Therefore, this discriminative classifier is less sensitive to class imbalance than generative classifiers, or other classification techniques, such as neural networks and decision trees (Zeng & Gao, 2009). E-SVM as a passive learning approach impose no extra training overhead and compensate the imbalance data in the optimization process. Compared to other cost-sensitive SVMs, like z-SVM that apply active learning for classification of imbalanced data (Imam *et al.*, 2006), the weights for classes are empirically determined during test mode. However, z-SVM requires more than one minority class sample to multiply the magnitude of positive support vectors by a particular small value of z estimated to bias the decision toward the majority negative class.

Since multiple representations can be generated from a single target to train these e-SVM classifiers, each classifier in an individual-specific ensemble is a different representation of an individual's face. Unlike similarity measurement methods, such as nearest neighbor schemes, e-SVMs do not necessarily compute distances to the other samples. Thus, combining e-SVMs into an ensemble may prevent over-fitting problems and simultaneously provides higher generalization performance (Li *et al.*, 2013a).

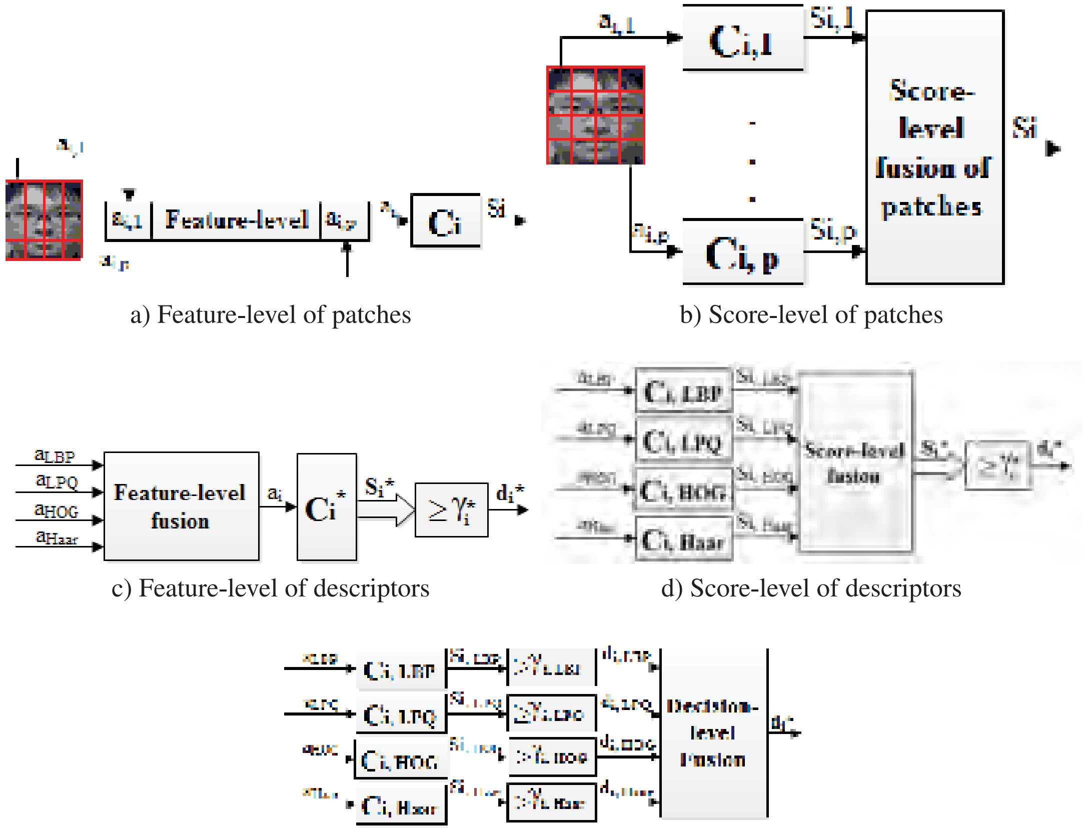
This method can be interpreted as an approach to sort non-targets by visual similarity to the individual, because estimated support vectors also belong to the non-targets. However, in this case, since each e-SVM is supposed to correctly classify only visually similar faces, these faces can be used as an additional target samples that can be employed either in calibrating decision boundary or defining decision thresholds. As another advantage of using e-SVMs, the support vectors can be exploited as the closest non-target samples to the single target reference for selection of the most similar non-targets. Setting different regularization parameters during training, produce different number of support vectors. These support vectors can be ranked and used to define decision thresholds, although it could be difficult due to inter-operability of cameras.

As an alternative, a pool of OC-SVM classifiers may be generated using ROIs of non-target individuals selected to provide accurate decision boundaries. The main difference of this approach with conventional one-class classification is that SVMs are trained based on non-target class samples rather than samples from class of interest. In this context, contrary to template matching (Bashbaghi *et al.*, 2014) that can be considered as a one-class classifier based on a single target reference still, OC-SVM can be defined as a method that either classifies non-target samples or rejects target samples during operation. Hence, the scores provided by OC-SVM classifiers can determine whether the input ROI patterns belong to non-target individuals or not and consequently target individuals are correctly detected.

3.1.2 Operational Phase

In the proposed system, different fusion approaches are applied to the proposed ensemble-based framework to achieve a higher level of generalization, and robustness (Connaughton *et al.*, 2013). Fusion techniques in such systems can be described as: (a) feature-level that aims to combine all the features extracted among patches into one feature vector in the feature space, (b) score-level attempts to combine the scores generated among patches using multiple classifiers trained per each patch, (c) feature-level concatenates several representations (descriptors), (d) score-level fusion of representations within the ensemble to provide the final score, and finally (e) decision-level of descriptors to produce the final response after applying decision thresholds as represented in Figure 3.3.

With the feature-level fusion of patches, features extracted from the patches isolated within the ROI are concatenated to construct a long feature vector that is dimensionally equivalent to the number of patches multiplied by the dimensionality of feature extraction techniques. PCA is applied to project data in such that features may be ranked according to covariance, and the most correlated features may be reduced, where only one SVM classifiers is subsequently trained per ROI. In score-level fusion of patches, a separate SVM classifier is trained on the features extracted from each patch, so that a number of classifiers identical to the number of patches is trained per ROI. Moreover, multiple representations are concatenated after applying PCA and then a single classifier is trained to perform feature-level fusion of descriptors. Scores are combined among multiple classifiers within the ensemble using the average function. Finally, decision-level fusion of descriptors consists in defining local decision thresholds for each descriptor specifically and exploiting majority vote to integrate their local decisions and produce the final decision. Decisions thresholds are defined using cumulative probability distribution function of non-target scores distribution at certain operating point of FPR=1% (Bashbaghi *et al.*, 2014).



a) Feature-level of patches

b) Score-level of patches

c) Feature-level of descriptors

d) Score-level of descriptors

e) Decision-level of descriptors

Figure 3.3 Five approaches for fusion of responses after extracting features from multiple patches and descriptors of an individual j (for $j = 1, 2, ..., N$) considered in this chapter.

3.1.2.1 Dynamic Classifiers Selection

In contrast to static approaches, the most competent classifiers in an individual's pool of classifiers that are trained over multiple face patches and representations can be selected and combined dynamically during operation in response to each probe ROI. Dynamic selection is used to improve the recognition accuracy by selecting the most competent classifiers and also to alleviate the computational cost. Hence, a novel approach is proposed to provide the best separation w.r.t. non-target samples in order to select an ensemble of classifiers based on a single high-quality target face still and many non-target low-quality video faces. Thus, the key idea is to allow the system to select those classifiers (face representations) that most properly discriminate target versus non-targets. In addition, this approach can improve the run-time speed in such applications by combining the selected classifiers rather than the entire pool. The proposed classifier selection method is formalized in Algorithm 3.1.

Algorithm 3.1 Dynamic ensemble selection method for individual j.

```
Input: Pool of diverse classifiers $P_j = \{c_{j,1}, \dots, c_{j,M}\}$, set of support vectors $\{SV_j\}$,
reference target still $G_j$, and the dataset of probe video ROIs $D_{test}$
Output: the set of the most competent classifiers $\{C^*\}$ for testing sample $t$ in $D_{test}$
for each probe ROI $t$ in $D_{test}$ do
    Divide $t$ into uniform patches $p$
    $\mathbf{a}_{i,p} \leftarrow$ extract ROI pattern $i$ from each patch $p$
    for each target individual $j$ do
        Project $\mathbf{a}_{i,p}$ into the feature space of $\{SV_j\}$ in $P_j$ and the target still $G_j$
        for each classifier $c_{j,k}$ in $P_j$ do
            if $Dist\left(\mathbf{a}_{i,p}, \mathbf{T}_{i,p}\right) \leq Dist\left(\frac{\sum_{s=1}^{s=|SV|} \mathbf{a}_{i,p}, \mathbf{V}_{i,p}^{s}}{|SV|}\right)$ then
                $\{C^*\} \leftarrow c_{j,k}$
            end if
        end for
        if $\{C^*\}$ is empty then
            Combine all classifiers $C$ in the pool to classify $t$ using mean function
        else
            Combine $\{C^*\}$ to classify $t$ using mean function
        end if
    end for
end for
```

The selection criteria (level of competence) per given ROI pattern has two components: (1) distance from non-target support vectors, and (2) closeness to the target reference still, where if the distance between the input pattern and the target still is lower than the distance from support vectors (average distances from all support vectors), then those classifiers are selected dynamically. Contrarily to the conventional approaches that use local neighborhood accuracy for measuring the level of competence, it is not necessary in this approach to define neighborhood by measuring the distance from all the validation data. However, Euclidean distance is employed to measure the distances between the input pattern and either target still or non-target support vectors.

3.1.2.2 Spatio-Temporal Fusion

In the proposed system, the head-face tracks are also exploited allowing for accumulation of scores associated with a same person to fulfill a robust spatio-temporal recognition. ROI captures for different individuals are regrouped into facial trajectories. Predictions for each individual are accumulated over time and if positive predictions surpass the detection threshold, then an individual of interest is detected. In particular, decision fusion module accumulates the ensemble scores S_j^* (obtained using score-level fusion) of each individual-specific ensemble over a fixed size window W according to:

$$d_j^* = \sum_{w=0}^{W-1} S_j^* \left[S_{i,p(W-w)} \right] \in [0,W] \tag{3.1}$$

3.2 Experimental Results and Discussions

Different aspects of the proposed framework are evaluated experimentally using Chokepoint (Wong *et al.*, 2011) and COX-S2V (Huang *et al.*, 2013a) still-to-video datasets. First, experiments assess the performance of classifiers trained on ROI patterns extracted using different feature extraction techniques. Second, experiments investigate the impact of patch configurations on the performance. Third, the performance of different levels and types fusion are compared. Finally, experiments show the effect of employing a tracker to form facial trajectories accumulate the ensemble predictions over consecutive frames in a trajectory and performing spatio-temporal recognition.

The size of the reference stills and captured ROIs are scaled to 48x48 pixels due to operational time. Libsvm library (Chang & Lin, 2011) is used in order to train e-SVMs and OC-SVMs. The same regularization parameters $C_1 = 1$ and $C_2 = 0.01$ are considered for all exemplars (w of a target sample is 100 times greater than non-targets). Previous study (Zhang & Wang, 2013) and experiments confirm that the optimal results will be achieved by choosing the misclassification costs (C_1 and C_2) based on the imbalance ratio. Differences greater than this will not improve the performance and on the other hand, the differences lower than this, may find worse decision boundary and degrade the performance.

3.2.1 Experimental Protocol

Ensemble of TMs (Bashbaghi *et al.*, 2014), ensemble of OC-SVMs, SVDL (Yang *et al.*, 2013), and ESRC-DA (Nourbakhsh *et al.*, 2016) are considered as the baseline and state-of-the-art FR systems to validate the proposed framework. In kNN experiments, eigenfaces of ROIs (Zhang *et al.*, 1997) are employed to compute the specialized kNN adapted for VS (VSkNN) based on k equals to 3 (1 target still and 2 nearest non-targets captured from background model) (Pagano *et al.*, 2014). To that end, the distance of the probe face are calculated from the target watch-list still along with the distance from the 2 nearest non-target captures from the training set. Thus, VSkNN score S_{VSkNN} is obtained as follows:

$$S_{VSkNN} = \frac{d_T}{d_T + d_{NT_1} + d_{NT_2}} \tag{3.2}$$

where d_T is the distance of the probe face from the target still, d_{NT_1} and d_{NT_2} are the distances from the nearest non-target captures, respectively.

Libsvm is also exploited in order to train OC-SVMs, where the regularization parameter n sets to 0.01 that indicates 1% of the non-target training data can be considered as support vectors. In SVDL experiment, 5 high-quality stills belonging to individuals of interest are considered as a gallery set and low-quality videos of non-target individuals are employed as a generic training set to learn a sparse variation dictionary. Three regularization parameters λ_1, λ_2, and λ_3 set to 0.001, 0.01, and 0.0001, respectively according to the default values defined in SVDL. The number of dictionary atoms are initialized to 80 based on the number of stills in the gallery set, where it is a trade-off between the computational complexity and the level of sparsity.

3.2.2 Results and Discussion

The performance of different aspects of the proposed framework using different feature extraction techniques with feature- and score-level fusion among patches is shown in Table 3.2 and Table 3.3 for Chokepoint and COX-S2V videos. Experiments are provided for non-overlapping patch configurations with 1, 4, 9, and 16 blocks (48x48, 24x24, 16x16, and 12x12 pixels, respectively). The scores of SVM classifiers trained over each patch are combined to provide the final score for each representation using score averaging. Noted that dimension of the representations vary, for instance the dimension of HOG and Haar depends on the resolution of the image and they typically produce a longer feature vector. Due to complexity and to avoid over-fitting, the number of dimensions are also reduced using PCA. An example of the ROC and inverted-PR curves obtained using ensemble of e-SVMs (4 blocks and HOG descriptor) is shown in Figure 3.4 with P1E_S1_C1 videos of Chokepoint.

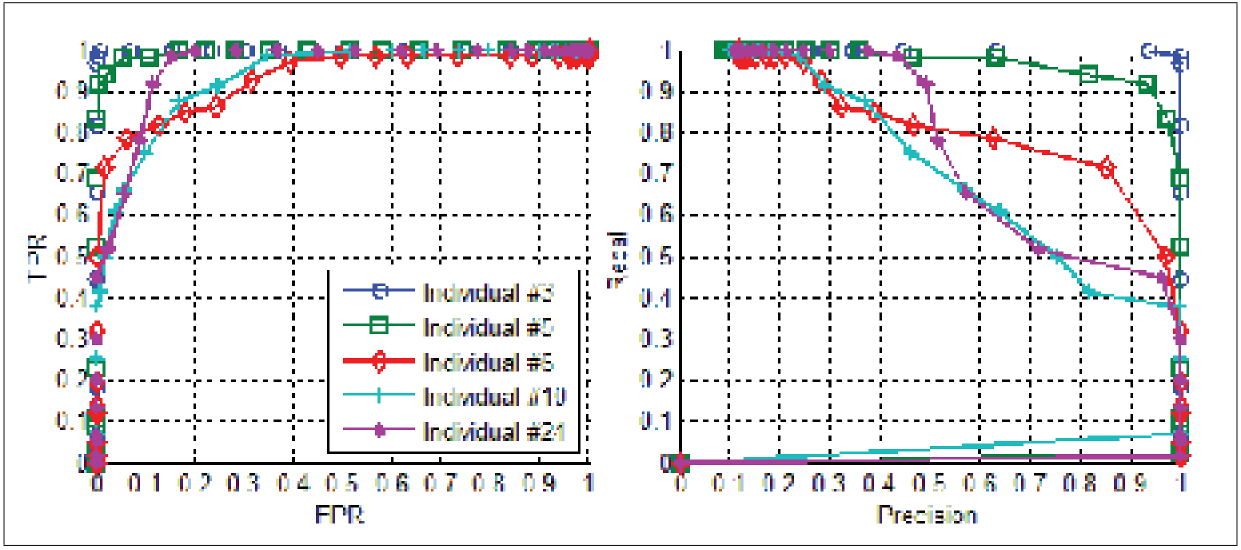


Figure 3.4 ROC and inverted-PR curves for a randomly selected watch-list of 5 individuals with Chokepoint video P1E_S1_C1.

The average values of pAUC(20%) and AUPR along with standard errors are presented in the Table 3.2 and Table 3.3 for different patch configurations.

Table 3.2 Average pAUC(20%) and AUPR accuracy of proposed systems at the transaction-level using feature extraction techniques (w/o patches) and videos of the Chokepoint dataset.

ROI - Patch Configurations	**Face Representations**							
	LBP (59 features)		**LPQ (256 features)**		**HOG (500 features)**		**Haar (2304 features)**	
	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR
1 block (48x48 pixels)								
All features	77.86±2.53	72.12±7.18	83.60±2.72	79.98±6.84	91.50±2.30	88.46±4.18	78.20±2.62	76.56±8.16
PCA features (max 64)	77.86±2.53	72.12±7.18	77.93±1.80	69.13±7.10	86.08±1.70	81.71±6.34	71.12±3.08	67.54±8.92
4 blocks (24x24 pixels)								
Score-level	79.53±2.34	74.71±8.76	79.20±2.66	76.65±8.40	91.03±0.84	88.02±4.32	84.41±2.38	81.82±7.42
Feature-level	78.00±2.76	75.16±6.50	80.00±2.46	76.24±4.28	79.50±3.10	77.36±7.18	72.44±2.68	69.80±4.00
9 blocks (16x16 pixels)								
Score-level	81.68±2.04	77.38±6.37	85.03±1.12	82.18±6.90	98.44±0.78	96.64±2.12	82.50±1.16	80.46±6.20
Feature-level	51.70±2.82	48.92±6.14	80.90±3.22	79.14±7.72	77.60±2.24	74.38±4.24	80.00±3.06	77.62±4.68
16 blocks (12x12 pixels)								
Score-level	33.60±2.32	32.78±2.82	52.70±2.24	49.70±4.42	65.30±3.04	61.12±6.62	70.00±2.40	68.82±7.28
Feature-level	30.50±1.24	28.82±6.00	35.00±2.40	32.78±4.96	71.10±3.54	69.78±4.16	70.56±3.38	67.28±4.94

As shown in Table 3.2, using patch-based method with 4, and 9 blocks (24x24 and 16x16 pixels, respectively) outperforms cases without patches (1 block). Patches with 16x16 pixels significantly outperforms case with large patches, and HOG in most cases provides better performance, especially when 9 blocks are used. The performance obtained using the smaller patches (12x12 pixels) is substantially lower, because features extracted from these small sub-images are not discriminant enough to generate robust classifier ensembles.

Feature-level fusion is also performed, where features extracted from patches are concatenated into a long feature vector and only one classifier is trained per each ROI representation. To reduce complexity, the dimension of features extracted from each patch is first reduced using PCA and then they are concatenated into the higher dimensional vector (for PCA projection, the 64 first eigenvectors are selected as features for LPQ, HOG and Haar descriptors). Concatenating features from larger blocks mostly provides higher performance. Longer ROI patterns obtained from more patches with smaller size may not perform well due to less of discriminative eigenvectors after applying PCA. However, training a separate classifier for each patch and combining local SVMs at a score-level typically achieves better performance, compared to training one global SVM based on the concatenated features extracted from all of the patches (feature-level fusion).

The experiments conducted over COX-S2V videos (Table 3.3) also suggest that the score-level fusion of patches can yield a better performance in contrast to the feature-level fusion of patches, due to encoding the pixels within each local patch into a different classifier separately.

Table 3.3 Average pAUC(20%) and AUPR accuracy of proposed systems at the transaction-level using feature extraction techniques (w/o patches) and videos of the COX-S2V dataset.

ROI - Patch Configurations	**Face Representations**							
	LBP (59 features)		**LPQ (256 features)**		**HOG (500 features)**		**Haar (2304 features)**	
	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR
1 block (48x48 pixels)								
All features	85.86±0.64	75.03±0.89	91.31±0.65	76.08±1.24	97.95±0.70	77.54±1.54	97.46±0.50	76.08±1.24
PCA features (max 64)	85.86±0.64	75.03±0.89	91.03±1.12	79.04±1.52	91.31±1.30	75.12±3.02	89.54±1.94	72.53±1.97
4 blocks (24x24 pixels)								
Score-level	92.31±1.14	80.38±1.30	95.40±1.14	84.26±1.46	98.47±1.58	86.70±1.80	97.70±0.47	80.73±1.72
Feature-level	84.59±1.10	73.00±1.58	89.64±0.26	82.00±0.73	88.08±2.30	68.30±1.86	89.90±1.32	78.04±0.89
9 blocks (16x16 pixels)								
Score-level	96.88±1.78	82.15±1.81	94.13±0.48	85.72±0.70	98.37±0.65	87.35±1.02	97.08±0.44	77.40±1.60
Feature-level	74.78±2.28	55.38±2.43	87.12±0.59	77.68±0.97	86.80±2.35	63.07±2.30	89.42±1.25	73.16±0.80
16 blocks (12x12 pixels)								
Score-level	76.10±2.28	49.98±3.37	86.62±0.82	75.60±0.92	92.95±1.06	80.01±1.46	92.96±0.38	76.77±1.42
Feature-level	69.93±1.05	48.44±3.85	80.72±1.05	69.98±0.74	91.64±1.47	75.02±2.14	93.68±0.48	64.63±2.15

Since each feature extraction technique performs inconstantly, applying fusion among them with dynamic classifier selection can provide higher level of performance. Table 3.4 presents a performance comparison for ensemble of classifiers designed with e-SVMs and OC-SVMs using feature- and score-level fusion of descriptors. The results of the proposed framework are also compared against baseline and state-of-the-art systems: VSkNN (Pagano *et al.*, 2014), SVDL (Yang *et al.*, 2013), ESRC-DA (Nourbakhsh *et al.*, 2016), and ensemble of TMs (Bashbaghi *et al.*, 2014) with the Chokepoint data. The performance achieved by combining the descriptors within static and dynamic ensembles at feature-level (concatenation) and score-level (mean function).

Using fusion of descriptors within the ensemble significantly improves performance over individual feature extraction techniques either with or without patches at transaction-level. Results indicate that the score-level fusion outperforms feature-level fusion, and 1 block (48x48 pixels) performs worse than other patch configurations. Feature-level fusion provides lower performance due to the effects of dimension of the concatenated vectors and the training of only one global SVM classifier. Accordingly, accurate local SVM classifiers leads to a robust ensembles for face screening, where patches size 16x16 pixels performs slightly better.

It can be seen from Table 3.4 that ensemble of e-SVMs outperforms ensemble of OC-SVMs, ensemble of TMs, VSkNN, SVDL and ESRC-DA. Performance of the FR system using VSkNN and SVDL is poor, mostly because of the significant differences between quality and appearances of the target face stills in the gallery set and video faces in the generic training set, as well as, imbalance of target versus non-target individuals observed during operation. It is worth noting that both VSkNN and SVDL are more suitable for close-set FR problems, such as face identification. Since each faces captured should be assigned to one of the target still in the gallery, therefore, many false positives occur. Moreover, SVDL can only apply as a complex global N-class classifier in contrast to the proposed ensemble of SVMs, due to sparse optimization and classification during operational phase.

Results indicate that, OC-SVM classifiers cannot classify target ROI patterns as discriminantly as e-SVM classifiers, because the target reference is not considered during training. Since the model (decision boundary) learned by OC-SVM is only based on low-quality non-target ROIs, and the quality of probe target ROIs are also similar to the training data, this model may fail to classify target ROIs precisely. In terms of number of blocks, ensemble of OC-SVMs using 9 blocks provides higher performance than others at score-level fusion. The proposed dynamic ensemble selection method is also assessed using 4, 9, and 16 blocks. The bottom of Table 3.4 shows that dynamic selection can improve accuracy and efficiency during operation by combining a lower number of classifiers. It slightly provides better results, where basically the larger the number of classifiers in the pool, the better the results achieved.

To validate the results, the aforementioned experiments are also repeated using the challenging COX-S2V dataset, where only 25 ROIs of target individual captured during operation are

Table 3.4 Average pAUC(20%) and AUPR performance of different implementations of the proposed framework at the transaction-level over Chokepoint videos. Results are shown using feature-, score-level fusion of patches and descriptors against reference state-of-the-art systems.

FR Systems		pAUC(20%)	AUPR
VSkNN (Pagano *et al.*, 2014)		19.00±0.40	16.48±0.90
SVDL (Yang *et al.*, 2013)		74.91±4.03	65.09±4.82
ESRC-DA (Nourbakhsh *et al.*, 2016)		97.16±1.28	76.97±6.73
Ensemble of TMs (Bashbaghi *et al.*, 2014)		85.60±1.04	82.78±7.06
Ensemble of OC-SVMs **1 block (48x48 pixels)**	**Feature-level**	71.34±5.78	64.07±5.96
	Score-level	86.10±1.06	81.62±7.82
4 blocks (24x24 pixels)	**Feature-level**	86.24±0.45	84.48±0.61
	Score-level	89.40±2.42	88.02±6.20
9 blocks (16x16 pixels)	**Feature-level**	96.73±0.43	91.55±4.43
	Score-level	97.40±0.40	95.72±2.64
16 blocks (12x12 pixels)	**Feature-level**	86.15±1.92	83.80±4.05
	Score-level	88.20±1.10	84.66±2.92
Dynamic Ensemble of OC-SVMs **4 blocks (24x24 pixels)**	**Score-level**	98.10±0.48	96.14±0.76
9 blocks (16x16 pixels)	**Score-level**	98.42±0.86	96.47±1.24
16 blocks (12x12 pixels)	**Score-level**	95.43±0.66	92.96±1.75
Ensemble of e-SVMs **1 block (48x48 pixels)**	**Feature-level**	92.90±0.82	90.20±5.06
	Score-level	92.28±0.54	90.95±2.84
4 blocks (24x24 pixels)	**Feature-level**	94.40±0.74	91.98±5.52
	Score-level	98.58±0.40	97.34±1.82
9 blocks (16x16 pixels)	**Feature-level**	89.80±0.12	89.24±0.44
	Score-level	100±0.00	99.24±0.38
16 blocks (12x12 pixels)	**Feature-level**	88.40±0.70	86.44±3.60
	Score-level	95.30±0.92	93.86±2.28
Dynamic Ensemble of e-SVMs **4 blocks (24x24 pixels)**	**Score-level**	100±0.00	98.86±0.90
9 blocks (16x16 pixels)	**Score-level**	100±0.00	99.31±0.46
16 blocks (12x12 pixels)	**Score-level**	97.71±1.06	94.60±3.12

matched against 2500 ROIs of non-target individuals. Results compare the state-of-the-art and baseline systems, and dynamic selection of OC-SVM and e-SVM classifiers. Table 3.5 presents the average transaction-level performance of systems over the COX-S2V data.

The results observed from Table 3.5 also confirm that ensembles of e-SVMs yield the more promising performance. Results are convincing even with high ratio of imbalances during

Table 3.5 Average pAUC(20%) and AUPR performance of different implementations of the proposed framework at the transaction-level over COX-S2V videos. Results are shown using feature-, score-level fusion of patches and descriptors against reference state-of-the-art systems.

FR Systems		pAUC(20%)	AUPR
VSkNN (Pagano *et al.*, 2014)		56.80±4.02	26.68±3.58
SVDL (Yang *et al.*, 2013)		69.93±5.67	44.09±6.29
ESRC-DA (Nourbakhsh *et al.*, 2016)		99.00±1.13	63.21±4.56
Ensemble of TMs (Bashbaghi *et al.*, 2014)		84.00±0.86	73.36±9.82
Ensemble of OC-SVMs **1 block (48x48 pixels)**	**Feature-level**	82.98±0.98	71.66±0.96
	Score-level	89.58±1.40	77.76±1.36
4 blocks (24x24 pixels)	**Feature-level**	84.94±1.13	75.84±1.62
	Score-level	90.04±0.88	82.61±0.68
9 blocks (16x16 pixels)	**Feature-level**	88.54±0.60	76.62±1.02
	Score-level	91.10±2.20	80.82±5.94
16 blocks (12x12 pixels)	**Feature-level**	83.91±0.83	74.94±1.26
	Score-level	89.28±1.44	79.96±1.08
Dynamic Ensemble of OC-SVMs **4 blocks (24x24 pixels)**	**Score-level**	94.00±1.78	86.72±1.94
9 blocks (16x16 pixels)	**Score-level**	95.78±0.52	87.48±4.06
16 blocks (12x12 pixels)	**Score-level**	95.58±1.15	87.65±1.72
Ensemble of e-SVMs **1 block (48x48 pixels)**	**Feature-level**	89.94±0.29	84.32±1.30
	Score-level	97.95±0.70	87.54±1.54
4 blocks (24x24 pixels)	**Feature-level**	91.12±1.18	83.24±1.56
	Score-level	99.74±0.06	90.21±0.56
9 blocks (16x16 pixels)	**Feature-level**	99.38±0.26	88.32±1.07
	Score-level	100±0.00	91.20±1.52
16 blocks (12x12 pixels)	**Feature-level**	96.62±0.76	80.60±1.50
	Score-level	98.47±0.32	87.48±1.02
Dynamic Ensemble of e-SVMs **4 blocks (24x24 pixels)**	**Score-level**	100±0.00	92.01±0.92
9 blocks (16x16 pixels)	**Score-level**	100±0.00	92.94±1.96
16 blocks (12x12 pixels)	**Score-level**	99.99±0.01	89.28±2.14

operation. The dynamic classifier selection method improves the performance and can provide higher accuracy for both OC-SVM and e-SVM ensembles.

To estimate the performance of different fusion approaches at a certain point of FPR, specific decision thresholds are applied to achieve desired FPR values. As illustrated in Figure 3.3 (c-e), only one threshold is defined for either feature- or score-level fusion, while decision thresholds dedicated to decision-level fusion (see Figure 3.3 (e)) are determined separately for each descriptor and the global decision is achieved through majority voting. The average performance of the system considering feature-, score-, and decision-level fusions at an exact point of FPR for both Chokepoint and COX-S2V datasets is presented in Table 3.6 and Table 3.7, respectively.

Table 3.6 Average performance of proposed system over Chokepoint videos at a certain point of FPR=1% using feature-, score, and decision-level fusion of descriptors within the ensemble.

Number of blocks	**Feature-level**			**Score-level**			**Decision-level**		
	TPR	FPR	F1	TPR	FPR	F1	TPR	FPR	F1
1 (48x48)	41.49±0.16	0.62±0.06	52.08±0.17	67.85±0.18	8.12±0.05	56.13±0.15	70.39±0.17	8.63±0.07	64.78±0.21
4 (24x24)	48.17±0.21	1.89±0.28	54.35±0.18	56.52±0.27	4.86±0.01	49.23±0.26	44.37±0.26	0.43±0.01	58.29±0.27
9 (16x16)	37.07±0.19	2.24±0.16	34.53±0.17	35.69±0.23	0.06±0.01	40.03±0.24	31.81±0.23	0.35±0.03	41.10±0.24
16 (12x12)	32.43±0.06	1.22±0.15	35.08±0.07	37.83±0.46	3.64±0.08	34.22±0.17	27.42±0.68	2.18±0.15	32.58±0.16

As shown in Table 3.6, decision-level fusion using 1 block (48x48 pixels) performs better than others in terms of F1 measures. Using 4 blocks (24x24 pixels) provides appropriate performance according to either F1 or TPR, retain FPR less than 1%. FPRs in feature- and score-level fusion are mostly greater than 1% due to inaccurate decision thresholds. Although defining decision thresholds per descriptor and using majority vote may lead to lower FPR, results after applying decision thresholds are generally poor. Defining decision thresholds in such an application may be a challenging task that affects overall system accuracy.

Table 3.7 Average performance of proposed system over COX-S2V videos at a certain point of FPR=1% using feature-, score, and decision-level fusion of descriptors within the ensemble.

Number of blocks	**Feature-level**			**Score-level**			**Decision-level**		
	TPR	FPR	F1	TPR	FPR	F1	TPR	FPR	F1
1 (48x48)	45.15±4.36	0.52±0.07	54.84±2.55	69.05±2.02	0.28±0.03	67.48±1.25	58.55±1.80	0.00±0.00	69.81±1.54
4 (24x24)	54.45±1.35	0.04±0.01	58.92±1.07	84.95±2.22	0.07±0.02	75.84±1.98	87.85±0.62	1.52±0.06	78.71±0.37
9 (16x16)	67.75±2.28	0.00±0.00	75.91±2.03	86.75±1.70	0.60±0.92	73.88±2.89	82.95±1.56	0.02±0.01	81.70±1.48
16 (12x12)	43.10±2.20	0.70±0.14	45.31±1.46	44.48±3.60	4.42±0.30	35.63±0.10	66.40±2.40	4.36±2.37	53.04±0.32

The results shown in Table 3.7 indicate that defining a dedicated threshold for each face descriptor at decision-level fusion using majority voting can achieve a higher accuracy in terms of F1 measure, mostly due to the lower values of FPR.

In another experiment, the proposed system is evaluated when a subset of background model is used during enrollment. To that end, videos captured from only one camera is considered to generate e-SVM ensembles and the system is tested on videos captured from other cameras. Table 3.8 presents the average performance of the proposed dynamic ensemble of e-SVMs using score-level fusion of descriptors with 9 blocks over COX-S2V videos, where for example videos captured from camera1 are used to train e-SVMs and the system is assessed on other videos captured from other cameras (camera2 to camera4).

Table 3.8 Average performance of the proposed system over COX-S2V videos, where a subset of background model is used for training.

Background model	**Video1**		**Video2**		**Video3**		**Video4**	
FR Systems	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR
Ensemble of e-SVMs	84.08±1.83	59.40±2.47	74.20±2.22	47.75±2.49	84.18±1.90	59.16±2.84	73.50±2.48	44.04±2.90
Dynamic Ensemble of e-SVMs	92.91±1.16	77.64±2.18	91.05±1.38	75.78±2.09	96.48±1.00	81.66±2.67	91.43±1.18	76.00±1.92

As shown in Table 3.8, considering a subset of background model during training e-SVMs can drastically reduce the performance in comparison with the results presented in Table 3.5. Since video2 and video4 are captured using a higher quality camera, better ensembles can be thus generated and subsequently, the performance of the system for other videos (video1 and video3) are relatively higher.

To analyze the impact of considering different number of unknown persons appearing in the operational scene, the number of unknown persons along with the target individual is varied from 100 to 300 and the AUPR performance is measured as displayed in Figure 3.5.

Since the proposed system is comprised of individual-specific ensembles that each one seeks to detect one target individual within the watch-list, as illustrated in Figure 3.5, it can perform consistently even with observation of severely imbalanced unknown persons during operation.

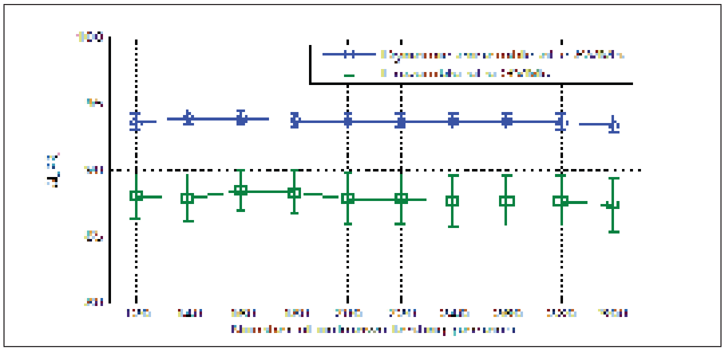


Figure 3.5 The analysis of system performance using different number of unknown persons during operation over COX-S2V.

The proposed ensemble of e-SVMs is also compared against ensemble of TMs as a baseline system at the trajectory-level. In this regards, the scores of individual-specific ensembles are gradually accumulated over a window of consecutive frames using a trajectory defined by the tracker. An example of accumulated scores over the trajectory is shown in Figure 3.6 and 3.7.

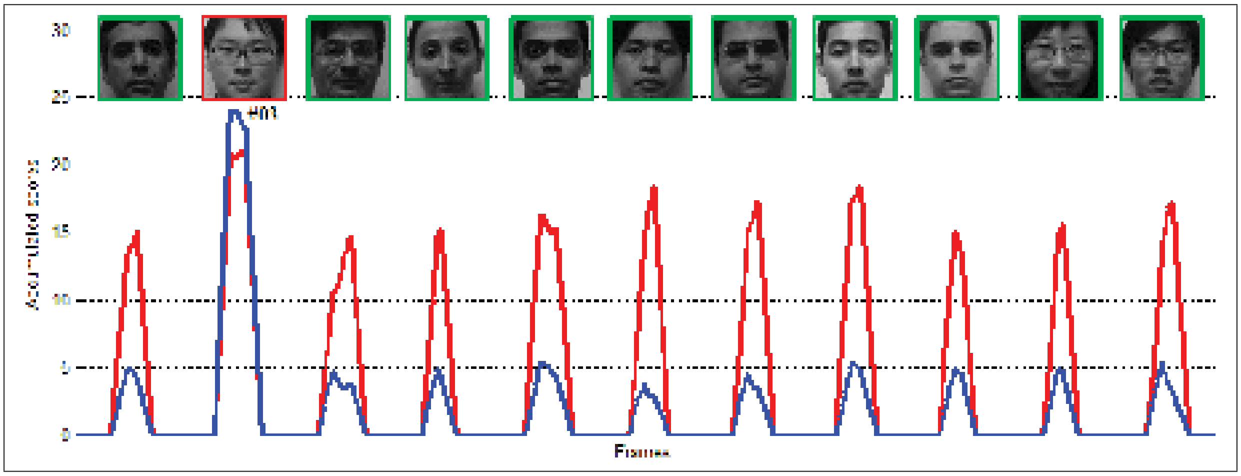


Figure 3.6 An example of the scores accumulated over windows of 30 frames with Chokepoint P1E_S1_C1 video using score-level fusion of descriptors with 4 blocks. Ensemble of e-SVMs is the blue curves and ensemble of TMs is the red curves.

As shown in Figure 3.6, the accumulated scores for target individual (ID#03) is significantly higher than all non-targets individuals for ensemble of e-SVMs, while the accumulated scores of non-targets are greater for ensemble of TMs. It can be observed that the accumulated scores of some non-target individuals are high, due to a higher number of false alarms. To assess the overall performance, the corresponding ROC curve may then plotted for each individual by varying the thresholds from 0 to 30 over accumulated scores, and the AUC are computed as overall performance of ensemble of e-SVMs. The average AUC for each watch-list individual across Chokepoint videos are provided in Table 3.9.

Table 3.9 AUC accuracy at the trajectory-level for ensemble of e-SVMs and TMs for a random selection of 5 watch-list individuals in the Chokepoint data.

	Individuals of interest					
	ID#03	**ID#05**	**ID#06**	**ID#10**	**ID#24**	**Average**
Ensemble of TMs (Bashbaghi *et al.*, 2014)	93.80±4.80	83.80±8.30	88.80±5.60	86.30±6.60	92.50±6.00	**89.04±6.26**
Ensemble of e-SVMs	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	**100.00±0.00**

Table 3.9 shows the average spatio-temporal recognition performance of the ensemble of e-SVMs is robust and higher than the baseline system.

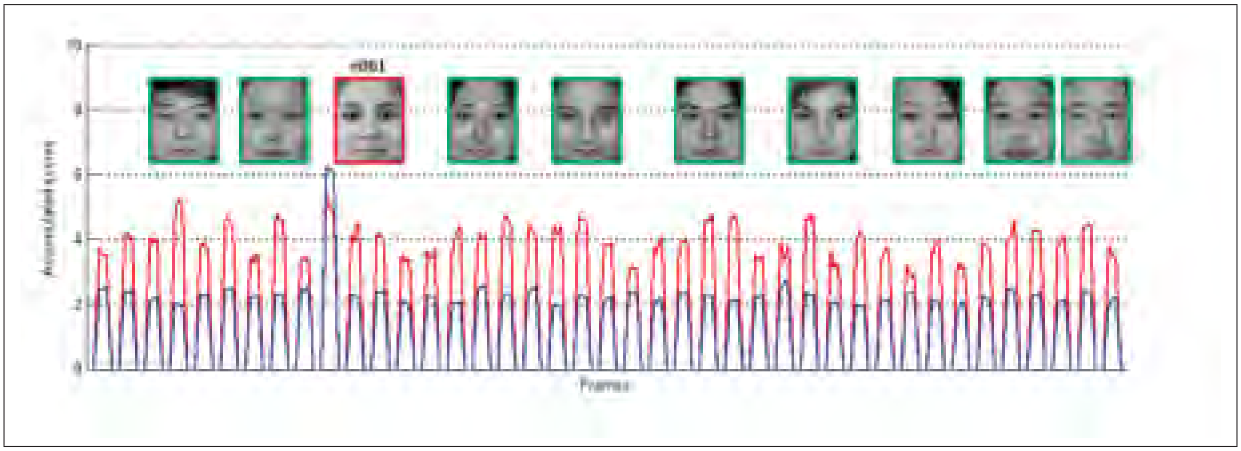


Figure 3.7 An example of the scores accumulated over windows of 10 frames with COX-S2V videos. Ensemble of e-SVMs is the blue curves and ensemble of TMs is the red curves.

It can be concluded from Figure 3.7 that ensemble of e-SVMs can outperform the baseline system under a severe imbalanced operational situation, where the target individual must be detected among more than a hundred people. Thus, the average spatio-temporal performance of the proposed and the baseline systems over COX-S2V videos are 100.00±0.00 and 86.01±2.36, respectively.

CHAPTER 4

DYNAMIC ENSEMBLES OF EXEMPLAR-SVMS FOR STILL-TO-VIDEO FACE RECOGNITION

In this chapter, an efficient and robust MCS is proposed for still-to-video FR. Multiple face representations and domain adaptation are exploited to generate an individual-specific ensemble of e-SVMs (Ee-SVM) per target individual using a mixture of facial ROIs captured in the enrollment domain (ED) (the single labeled high-quality still of target and cohort captured under controlled conditions) and the operational domain (OD) (i.e., an abundance of unlabeled facial trajectories captured by surveillance cameras during a calibration process). Facial models are adapted to the OD by training the Ee-SVMs using a single labeled target still ROI versus cohort still ROIs, along with unlabeled non-target video ROIs. Several training schemes are considered for DA of ensembles according to utilization of labeled ROIs in the ED and unlabeled ROIs in the OD.

During enrollment of a target individual, semi-random feature subspaces corresponding to different face patches and descriptors are employed to generate a diverse pool of classifiers that provides robustness against different perturbations frequently observed in real-world surveillance environments. In this chapter, two application scenarios are investigated to design the individual-specific ensembles. In the first scenario, a validation set is employed together with a global criterion (measuring the significance of each patch on the overall performance) in order to rank and select patches and subspaces. In contrast, a local distance-based criterion is used in the second scenario to rank subspaces without employing a validation set. In particular, various ranked feature subspaces are sampled from face patches represented using state-of-the-art face descriptors, instead of randomly sampling from the entire ROIs. Pruning of the less accurate classifiers is performed to store a compact pool of classifiers in order to alleviate computational complexity.

During operations, a subset of the most competent classifiers is dynamically selected/weighted and combined into an ensemble for each probe using a novel distance-based criteria. Internal criteria are defined in the e-SVM feature space that rely on the distances between the input probe to the target still and non-target support vectors. In addition, persons appearing in a scene are tracked over multiple frames, where matching scores of each individual are integrated over a facial trajectory (i.e., group of ROIs linked to the high-quality track) for robust spatio-temporal FR. The proposed system is efficient, since the criteria to perform DS and weighting allows to combine a lower restrained number of the most relevant classifiers within the individual-specific ensembles.

Videos from the COX-S2V (Huang *et al.*, 2013a) and Chokepoint (Wong *et al.*, 2011) datasets are employed to evaluate and compare the performance of the proposed system against state-of-the-art methods. These datasets contains a high-quality reference still from the ED and low-quality videos of individuals captured under uncontrolled conditions in different ODs. Experimental results are obtained at the transaction- and trajectory-levels in the ROC and precision-recall spaces. The results indicate that the proposed system provides state-of-the-art accuracy, yet with a significantly lower computational complexity. The contents of this chapter have been published in the journal of "Pattern Recognition" (Bashbaghi *et al.*, 2017a) and the International Conference in Pattern Recognition Application and Methods (Bashbaghi *et al.*, 2017c).

4.1 Dynamic Individual-Specific Ee-SVMs Through Domain Adaptation

A novel ensemble learning approach is proposed in this chapter to design accurate classification systems for each target individual enrolled to a still-to-video FR system. In particular, to improve robustness to intra-class variations, individual-specific Ee-SVMs models the single reference still ROI for the OD using several diverse e-SVMs based on multiple face representations and domain adaptation. During enrollment, each patch-wise e-SVM is trained for a different patch, descriptor and feature subset extracted from the single reference still ROI of the target individual (in the ED) versus those extracted from the abundance of still and video ROIs of non-target individuals (in either ED and OD). Several training schemes are proposed for unsupervised DA according to assumptions made for unlabeled video ROIs from the OD.

Two different scenarios are investigated for the design phase to select the most discriminant among a large number of representation subspaces (descriptors and feature subsets of a patch) for enrollment of target individuals (Ee-SVMs design). In the first design scenario, a validation set, containing stills and videos of some random non-target individuals, is exploited with a global criterion to effectively adapt the system to the actual context. Thus, the most accurate e-SVM classifiers (i.e., discriminative representation subspaces) are selected by ranking trained e-SVMs using a criterion based on the area under precision-recall curve (Cheplygina & Tax, 2011), where these subspaces are used for enrollment of a target individual. In the second design scenario, the most informative representation subspaces are selected without considering a validation set. A local distance-based criterion is applied to rank and prune them, where the best subspaces are selected for enrollment of a target individual.

Since capture conditions change over time, the best ensemble to recognize the target individual will vary according to the given probe ROI. Pre-selection of the most discriminative representation subspaces during the design phase, as well as, selecting or weighting the most competent classifiers during the operational phase can provide a higher level of performance at a lower computational complexity in such a real-time application, unlike employing fusion over the entire pool.

4.1.1 System Overview

A block diagram of the proposed MCS for still-to-video FR is shown in Figure 4.1. It generates a diverse and compact pool of classifiers during the design phase, and selection and weighting ensembles dynamically during the operational phase. Each step of the proposed system is described in the following subsections.

During the design phase (Enrollment/Design phase), a pool of diverse e-SVM classifiers is generated per individual of interest. Multiple different facial representations are produced over all patches for several face descriptors and random subspaces. The parameters of the proposed system, such as number of patches, number and size of feature subspaces are defined in this

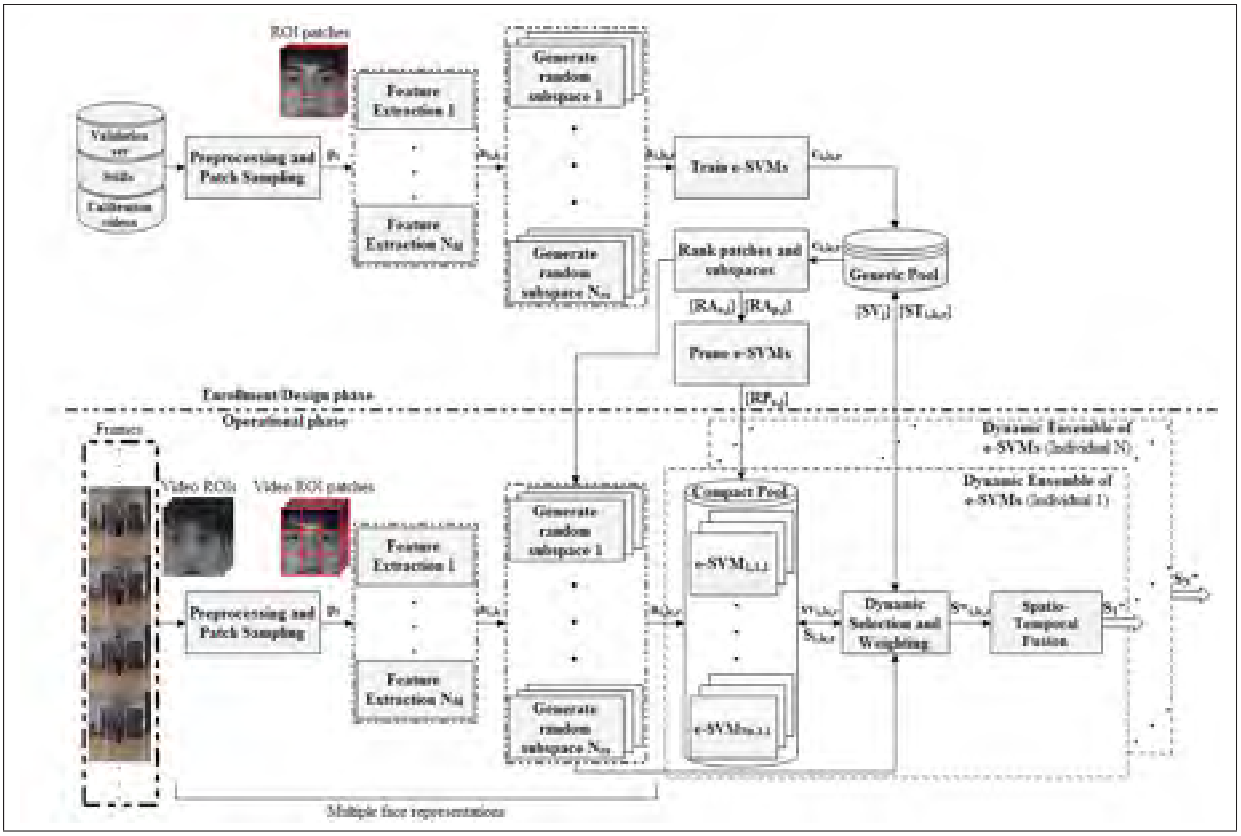


Figure 4.1 The enrollment and operational phases of the proposed multi-classifier system for accurate still-to-video FR.

phase. Different number of classifiers are trained for each patch based on their significances on performance using the best subspaces (representations) that were already ranked.

During the operational phase, classifiers of the pool are selected or weighted dynamically according to competence for classifying the given input probe (ROI), and then their scores are combined to obtain the final score. The proposed system exploits two levels of information fusion. First, the fusion of subspace-wise classifiers selected during operations from corresponding face descriptor (patch-level fusion), and then the fusion of patch-wise classifiers generated by the face descriptors (descriptor-level fusion).

4.1.2 Design Phase (First Scenario)

In this scenario for the design phase, a compact pool of e-SVM classifiers is generated using semi-random subspaces pruned based on the most informative pre-ranked patches. This phase is performed off-line, and as shown in Figure 4.1 (Enrollment/Design phase), it consists of patch-wise feature extraction, training patch-wise e-SVMs, as well as, ranking patches and pruning subspaces to select the best subspaces (representations). Note that in this scenario, the labeled stills and video trajectories correspond to some unknown individuals or actors appearing in the scene, and are used to estimate system parameters and pre-selection of the best subspaces. Then, the pre-selected subspaces are used to design an Ee-SVMs for individuals of interest based on a single labeled still.

The validation set D consists of labeled high-quality stills and unlabeled low-quality videos defined as $D = \left\{ST_1^l, \dots, ST_j^l, \dots, ST_{N_a}^l \cup T_1^l, \dots, T_j^l, \dots, T_{N_a}^l \cup T_1^u, \dots, T_v^u, \dots, T_{N_v}^u\right\}$, where ST_j^l and T_j^l represent the labeled still and video trajectory of individual j, respectively, and T_v^u denotes the unlabeled video trajectory of unknown person v. N_a indicates the number of unknown non-target individuals in the validation set, where the number of videos is equal to N_v. All the stills and videos are segmented and scaled to the resolution of M_cxN_c. As illustrated in Figure 4.1, all still ROIs of ST_j^l and video ROIs of T_j^l and T_v^u are first divided into m_cxn_c pixels patches $P_j^l = \left\{p_i^l\right\}$ and $P_v^u = \{p_i^u\}$, where $i = [1, 2, \dots, N_p]$ and $N_p = (M_c/m_c) \times (N_c/n_c)$ is the total number of patches. Afterwards, feature extraction techniques (face descriptors) $FD = \{f_k\}$ are applied to extract feature sets $F_j^l = \left\{\mathbf{a}_{i,k}^l\right\}$ and $F_v^u = \left\{\mathbf{a}_{i,k}^u\right\}$ from patch p_i, for $k = 1, 2, \dots, N_{fd}$ and N_{fd} is the number of face descriptors. Thus, $\mathbf{a}_{i,k}$ defines the descriptor f_k extracted from patch p_i. Then, different random subspaces $RS = \{s_r\}$ with the dimension N_d are randomly selected from F_j^l and F_v^u to generate random subspaces $R_j^l = \left\{\mathbf{a}_{i,k,r}^l\right\}$ and $R_v^u = \left\{\mathbf{a}_{i,k,r}^u\right\}$, for $r = 1, 2, \dots, N_{rs}$, and N_{rs} is the total number of random subspaces. Hence, $\mathbf{a}_{i,k,r}$ denotes the feature subspaces s_r randomly selected from $\mathbf{a}_{i,k}$.

To construct a compact pool of classifiers $P_c = \{Ej | 1 \leq j \leq N_a\}$, ensemble of e-SVM classifiers $E_j = \left\{C_l | 1 \leq l \leq N_P \cdot N_{fd} \cdot N_{rs}\right\}$ are trained to enroll a target individual j. The number of random subspaces $RP_{s,j} = \{s_r | 1 \leq r \leq N'_{rs}\}$ is determined based on the significance of patches $RA_{p,j}$ and their rankings $RA_{s,j}$ to train accurate classifiers $c_{i,k,r}$ (See Algorithm 4.3). However, all the subspaces $RS = \{s_r\}$ are employed to construct a generic pool of classifiers $P_g = \{Ej | 1 \leq j \leq N_a\}$, where $E_j = \left\{C_l | l = 1, 2, \ldots, N_P \cdot N_{fd} \cdot N_{rs}\right\}$ as formalized in Algorithm 4.1.

Algorithm 4.1 Generic pool generation.

```
Input: Validation set $D = \left\{ST_1^l, \ldots, ST_j^l, \ldots, ST_{N_a}^l \cup T_1^l, \ldots, T_j^l, \ldots, T_{N_a}^l \cup T_1^u, \ldots, T_v^u, \ldots, T_{N_v}^u\right\}$
Output: Generic pool of e-SVM classifiers $P_g = \{Ej | 1 \leq j \leq N_a\}$
{Constructing an ensemble of e-SVMs}
for each individual $j$ in $D$ do
    Divide $ST_j^l$, and $T_v^u$ into patches $P_j^l$ and $P_v^u$ of size $m_c$x$n_c$
    for each patch $i = 1...N_p$ do
        for each face descriptor $k = 1...N_{fd}$ do
            {Patch-wise feature extraction}
            $\mathbf{a}_{i,k} \leftarrow$ extract face descriptors $f_k$ from patch $p_i$
            for each random subspace $r = 1...N_{rs}$ do
                $\mathbf{a}_{i,k,r} \leftarrow$ randomly sample subspaces $s_r$ from $\mathbf{a}_{i,k}$
                {Training patch-wise e-SVM classifiers}
                $E_j \leftarrow$ train a classifier $c_{i,k,r}$
            end for
        end for
    end for
end for
```

As formulated in the Algorithm 4.1, labeled still ST_j^l and unlabeled video ROIs T_v^u in the validation set D are employed to train patch-wise e-SVM classifiers and subsequently, to build a generic pool of classifier $P_g = \{Ej | 1 \leq j \leq N_a\}$ based on DA using multiple face descriptors. To that end, an ensemble of e-SVMs E_j is constructed for each individual in D and stored within the generic pool.

Semi-random subspaces selected during this phase are utilized to increase the probability of generating representative facial models that are robust to nuisance factors existing in the surveillance environments. However, due to a loss of information in some of the subspaces, selecting a suitable size of patches and random subspaces are essential. The time complexity and accuracy are dependent to these parameters. Smaller rate of random sampling causes to perform faster, but simultaneously it may miss useful discriminant features subsets. On the other hand, larger rate may also cause less diversity among classifiers.

4.1.2.1 Patch-Wise Feature Extraction

In this chapter, the patches in each face are represented using LPQ and HOG descriptors (Ahonen *et al.*, 2008; Deniz *et al.*, 2011), although many other face descriptors may be suitable. The choice of face descriptors is based on the complementary robustness that they provide to the nuisance factors in surveillance environments (Bashbaghi *et al.*, 2014). Previous study suggests that the combination of these descriptors is capable of providing a high level of discrimination on the SSPP problem (Bashbaghi *et al.*, 2015, 2017a).

LPQ extract texture features of the face images from frequency domain through Fourier transform and has shown high robustness to motion blur. LPQ is based on the blur insensitive property of the Fourier phase spectrum. The phase is computed in local rectangular M-by$M$ neighborhoods N_x at each pixel position x of the image $f(x)$ using a short-term Fourier transform defined by:

$$F(\mathbf{u},x) = \sum_{y \in N_x} f(x-y)\, e^{-j2\pi \mathbf{u}^T y} = \mathbf{w_u}^T \mathbf{f_x} \tag{4.1}$$

where $\mathbf{w_u}$ is the basis vector of the 2-D discrete Fourier transform at frequency $\mathbf{u}$, and $\mathbf{f_x}$ is another vector containing all M^2 values of f in N_x. It is examined for all positions $x \in \{x_1, x_2, \ldots, x_N\}$ at four frequency points $\mathbf{u} \in \{\mathbf{u_1}, \ldots, \mathbf{u_4}\}$ that results in a vector $\mathbf{F_x}$. The phase information is obtained using the signs of each component in the $\mathbf{F_x}$ by a simple scalar quantizer $q_j(x)$, where $q_j(x)$ is the jth component of the Fourier coefficients. Then, the label image $f_{LPQ}(x)$ with blur invariant LPQ values is represented by eight binary coefficients $q_j(x)$ as integer values between 0-255 using the binary coding $f_{LPQ}(x) = \sum_{j=1}^{8} q_j(x) 2^{j-1}$. Finally, the histograms of labels $f_{LPQ}(x)$ from different non-overlapping rectangular regions are concatenated to build the 256-dimensional LPQ face descriptor.

On the other hand, HOG extracts gradients, and it is more robust to pose and scale changes, as well as, rotation and translation. In particular, the occurrences of gradient orientations are counted in each local neighborhood of an image. The image is divided into different blocks and cells (small connected regions) for a block spacing stride of l pixels. Then a histogram of gradient orientations is computed for each cell within the blocks. According to the sign of gradients, the channels of each histogram can be varied over $0-180°$ or $0-360°$ for unsigned and signed, respectively with 9 orientation bins. The histograms are normalized using color and Gamma correction with L2-Hys threshold for robustness against illumination and scale. Finally, the combination of normalized group of histograms in all cells and blocks represents the HOG face descriptor.

4.1.2.2 Training Patch-Wise E-SVM Classifiers

Designing accurate classifiers for a MCS under imbalanced data situation is a challenging issue (Bashbaghi *et al.*, 2015). SVM is a well-known and widely used discriminative classifier that finds the optimal hyperplane to separate data patterns into binary classes. Thus, specialized 2-class SVMs are used to generate a pool of classifiers. Conventional 2-class SVM classifiers typically fail to find an optimal decision boundary in case of imbalance data (Zeng & Gao, 2009). However, different error costs (DEC) method (Veropoulos *et al.*, 1999) can be used to assign two misclassification cost values C^+ and C^- to manipulate the SVM objective function as follows:

$$min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^2 + C^+ \sum_{[i|y_i=+1]}^{l} \xi_i + C^- \sum_{[i|y_i=-1]}^{l} \xi_i \tag{4.2}$$

where $\mathbf{w}$ is the weight vector, b is the bias term, C^+ and C^- are the positive and negative misclassification costs to control the weight, respectively.

In the specialized approach proposed according to the existing constraints, classifiers are trained using a single target reference stills against many non-target samples. A method called exemplar-

SVM (e-SVM) (Malisiewicz *et al.*, 2011a) has been proposed to train a separate SVM classifier with DEC for each individual of interest. It has shown effectiveness and generalization to design an individual-specific ensembles for still-to-video FR, where diversity of an e-SVM pool is provided using multiple representations (Bashbaghi *et al.*, 2015). It is worth mentioning that training many different e-SVM classifiers based on multiple representations and then combining their scores may avoid the issue of over-fitting. Since there is only a single positive sample in the training set, its error should be weighted much higher than the negative samples to avoid the skewness toward negatives. Let **a** be the target ROI pattern, **x** and U be sets of non-target ROI patterns (either labeled still ROIs or unlabeled video ROIs depending on the different training schemes) and their number, respectively. The cost function of e-SVM using a linear kernel is formalized as follows:

$$min_{\mathbf{w},b}\mathbf{w}^2+C_1\max(0,1-\left(\mathbf{w}^T\mathbf{a}+b\right)+C_2\sum_{\mathbf{x}\in U}\max\left(0,1-\left(\mathbf{w}^T\mathbf{x}+b\right)\right) \tag{4.3}$$

where C_1 and C_2 define the regularization weights, **w** is the classifiers weight vector, b is the bias.

In order to learn the individual-specific Ee-SVM for target individual j based on DA, the 5 training schemes have been considered by employing either labeled still ROIs ST_j^l from the cohort or other non-target individuals or unlabeled video ROIs T_v^u captured from the operational domain.

a. Scheme 1 (target still ROI vs non-target still ROIs): The single labeled target still and non-target still ROIs from cohort model are employed to train e-SVMs without exploiting unlabeled video ROIs. Thus, videos in the OD are not employed for DA (see Figure 4.2 (a)).

b. Scheme 2 (target still ROI vs non-target video ROIs): The single labeled target still ROI are considered with an abundance of unlabeled non-target video ROIs from the OD (see Figure 4.2 (b)).

c. Scheme 3 (target still ROI vs non-target stills and video ROIs): Labeled non-target still ROIs from the cohort model are considered in addition to video ROIs from the OD (see Figure 4.2 (c)).

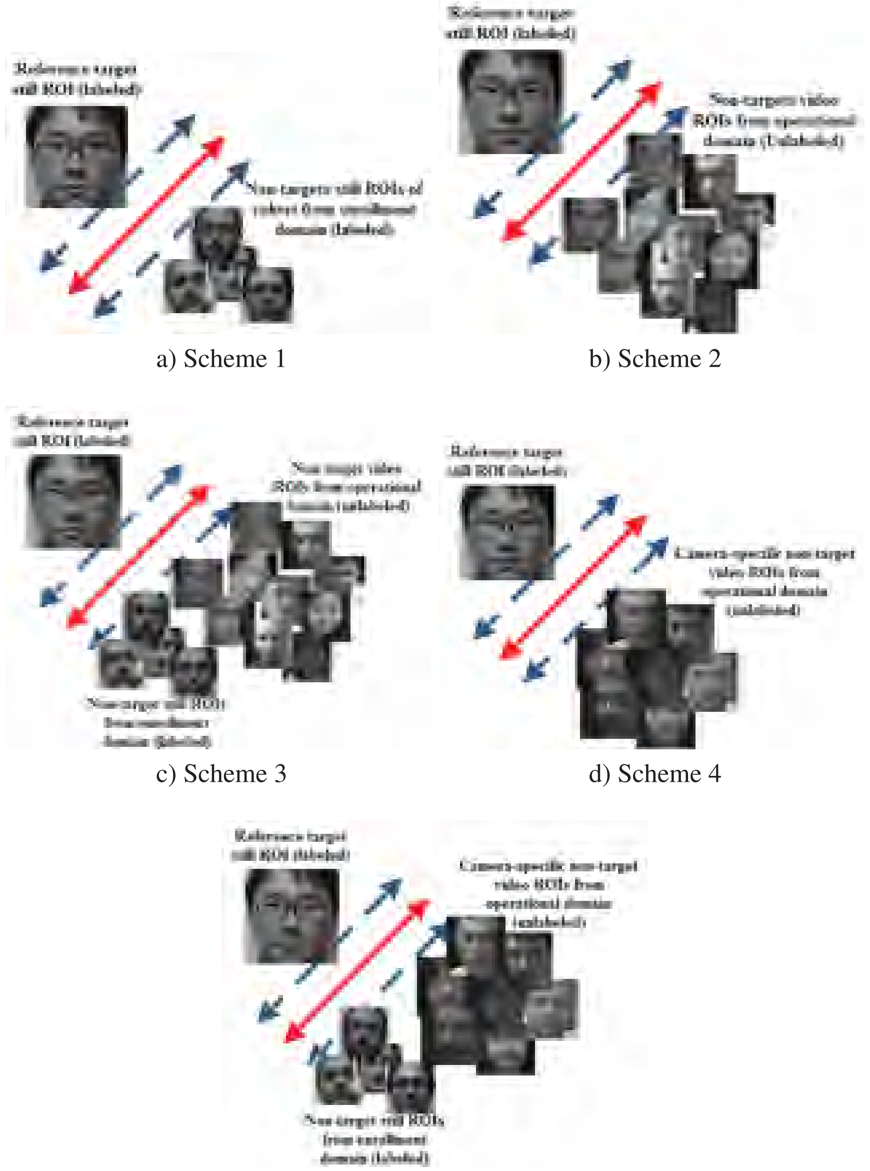
d. Scheme 4 (target still ROI vs unlabeled non-target camera-specific video ROIs): Unlabeled video ROIs captured using a specific camera FoV are exploited along with the labeled target still ROI in order to construct a camera-specific pool. Thus, several camera-specific pools equivalent to the number of surveillance cameras are constituted (see Figure 4.2 (d)).

e. Scheme 5 (target still vs non-target stills and camera-specific video ROIs): Labeled non-target still ROIs with unlabeled camera-specific video ROIs are considered versus the single target still ROI in order to build several camera-specific pools (see Figure 4.2 (e)).

To assess the 5 aforementioned training schemes, all the classifiers in the generic pool are tested to obtain the system performance. However, the best scheme is adopted to learn the individual-specific Ee-SVMs in the proposed system. To accomplish DA, unlabeled video ROIs captured from the OD allow to incorporate the knowledge of operational domain during generation of the pool. Therefore, an unsupervised DA approach is considered, where labeled still ROIs from the cohort model and unlabeled video ROIs captured from the OD are employed to train classifiers in the enrollment domain. As illustrated in Figure 4.2 (c), this training scheme favors the transfer of knowledge from either ED or OD to the classifiers trained specifically for each individual of interest.

4.1.2.3 Ranking Patch-Wise and Subspace-Wise e-SVMs

During the design prior to the enrollment, $N_p{\cdot}N_{rs}$ classifiers are trained for individuals in the validation set according to each face descriptor f_k. Then, these classifiers are combined using the mean fusion function over the random subspaces s_r (patch-level fusion). Subsequently, N_p classifiers are evaluated and ranked $RA_{p,j}$ using the global system performance based on the area under precision-recall (AUPR) as formulated in Algorithm 4.2. Noted that N_{rs} constant


Reference target still ROI (labeled)
Non-targets still ROIs of cohort from enrollment domain (labeled)
Reference target still ROI (labeled)
Non-targets video ROIs from operational domain (Unlabeled)
Reference target still ROI (labeled)
Non-target video ROIs from operational domain (unlabeled)
Non-target still ROIs from enrollment domain (labeled)
Reference target still ROI (labeled)
Camera-specific non-target video ROIs from operational domain (unlabeled)
Reference target still ROI (labeled)
Camera-specific non-target video ROIs from operational domain (unlabeled)
Non-target still ROIs from enrollment domain (labeled)

a) Scheme 1

b) Scheme 2

c) Scheme 3

d) Scheme 4

e) Scheme 5

Figure 4.2 A 2-D illustration of e-SVM in the feature space trained using different classification schemes according to DA. (a) a target still vs labeled non-target still ROIs of ED, (b) a target still vs unlabeled non-target video ROIs of OD, (c) a target still vs labeled non-target still ROIs of ED and video ROIs of OD, (d) a target still vs unlabeled non-target camera-specific video ROIs of OD, and (e) a target still vs labeled non-target still ROIs of ED and unlabeled non-target camera-Specific video ROIs of OD.

subspaces are selected from each patch, because it is tended to rank the significance of patches p_i based on the information encapsulated in each one.

In addition, to rank the subspaces s_r selected randomly from each patch p_i, the $N_p{\cdot}N_{rs}$ classifiers in the P_g are combined over the patches and the corresponding performance is similarly evaluated as in Algorithm 4.2. Thus, each feature subset is ranked and its corresponding classifier retained in $RA_{s,j}$ according to ranking of patches already preserved in $RA_{p,j}$.

Algorithm 4.2 Ranking of patch-wise and subspace-wise e-SVMS.

```
Input: Validation set $D$ and generic pool $P_g$
Output: Ranking of patches $RA_{p,j}$ and subspaces $RA_{s,j}$
for each individual $j$ in $D$ do
    for each face descriptor $k = 1...N_{fd}$ do
        {Ranking patch-wise classifiers}
        for each patch $i = 1...N_p$ do
            $RA_{p,j} \leftarrow \{\emptyset\}$
            Combine classifiers $c_{i,k}$ over random subspaces $s_r$ using the mean fusion function
            $RA_{p,j} \leftarrow$ rank patches $p_i$ in descending order of the AUPR obtained using $c_{i,k}$
        end for
        {Ranking subspace-wise classifiers}
        for each random subspace $r = 1...N_{rs}$ do
            $RA_{s,j} \leftarrow \{\emptyset\}$
            Combine classifiers $c_{k,r}$ over patches $p_i$ using the mean fusion function
            $RA_{s,j} \leftarrow$ rank subspaces $s_r$ in descending order of the AUPR obtained using $c_{k,r}$
        end for
    end for
end for
```

These ranking processes allow the pre-selection of e-SVM classifiers according to the best representations (feature subsets) during the design. It allows generating the less number of more accurate classifiers for each patch through patch ranking during the enrollment of target individuals.

4.1.2.4 Pruning Subspaces-Wise e-SVMs

After ranking patches and subspaces, a pruning process is used to select a variable numbers of the ranked subspaces from each patch as shown in Algorithm 4.3. A larger the number of subspaces are selected for the most relevant patches. In order to select different number of subspaces for each patch, a criterion is deployed as follows according to the overall AUPR performance obtained using all the classifiers in the pool $c_{i,k,r}$ and AUPR performance gained by corresponding all the classifiers of each patch $c_{i,k}$:

$$N'_{rs} = \left\lceil N_{rs}.\frac{AUPR\left(c_{i,k}\right)}{AUPR\left(c_{i,k,r}\right)} \right\rceil \tag{4.4}$$

where R_{pruned} contains N'_{rs} ranked subspaces s_r (integer values using a ceiling function) for each patch p_i. It allows to constitute the compact pool and accordingly, the dynamic classifier selection can be accomplished with the lowest number of classifiers during operations. However, the best subspaces are found during the design phase and those subspaces are employed to train e-SVMs for each individual in the watch-list during the enrollment phase.

Algorithm 4.3 Pruning subspace-wise e-SVMs and compact pool generation.

```
Input: Validation set $D$, generic pool $P_g$, ranked patches $R_{p,j}$, ranked subspaces $R_{s,j}$, and phase $phase$
Output: Compact pool of e-SVM classifiers $P_c = \{Ej | 1 \leq j \leq N_a\}$
for each individual of interest $j = 1...N_a$ do
    for each face descriptor $k = 1...N_{fd}$ do
        if design phase then
            for each patch $i = 1...N_p$ in the $R_{patch}$ do
                $N'_{rs} \leftarrow \left\lceil N_{rs}.\frac{AUPR(c_{i,k})}{AUPR(c_{i,k,r})} \right\rceil$
                $RP_{s,j} \leftarrow$ select $N'_{rs}$ subspaces from $R_{s,j}$ for each patch $p_i$
            end for
        end if
        if enrollment phase then
            for each random subspace $r = 1...N'_{rs}$ in the $RP_{s,j}$ do
                $E_j \leftarrow$ train $c_{i,k,r}$ to construct a compact pool of classifiers
            end for
        end if
    end for
end for
```

4.1.3 Design Phase (Second Scenario)

In this scenario relies on the over-produce and select paradigm, where a large number of subspaces are generated for each individual of interest during the design phase of the system. Then, e-SVM classifiers are trained and the best subspaces are selected during the enrollment phase. In the proposed system, several feature subspaces are randomly produced for each patch, and these subspace are ranked $RA_{s,j}$ based on a distance-based local criterion to select the best set of subspaces ($N'_{rs} \ll N_{rs}$). They can be employed to construct a compact pool of classifiers as presented in Algorithm 4.4.

Algorithm 4.4 Ranking subspace-wise e-SVMs and compact pool generation.

```
Input: Labeled still ROIs of target individuals $ST^l_1, \ldots, ST^l_j, \ldots, ST^l_{N_a}$ and unlabeled video ROIs of
  non-target individuals $T^u_1, \ldots, T^u_v, \ldots, T^u_{N_v}$, and phase $phase$
Output: Compact pool of e-SVM classifiers $P_c = \{Ej | 1 \leq j \leq N_a\}$
for each individual of interest $j = 1...N_a$ do
    for each patch $i = 1...N_p$ do
        for each face descriptor $k = 1...N_{fd}$ do
            if $phase = design$ then
                for each random subspaces $r = 1...N_{rs}$ do
                    $\mathbf{a}_{i,k,r} \leftarrow$ randomly sample subspaces $s_r$ from $\mathbf{a}_{i,k}$
                end for
            end if
            if $phase = enrollment$ then
                $E_j \leftarrow$ train a classifier $c_{i,k,r}$
                $RA_{s,j} \leftarrow$ rank subspaces in descending order based on the dist $\left(\mathbf{ST}_{i,k,r}, \mathbf{sv}_{i,k,r}\right)$
                {Constructing a compact pool (enrollment)}
                for random subspaces $r = 1...N'_{rs}$ in the $RA_{s,j}$ do
                    $E_j \leftarrow$ preserve $c_{i,k,r}$ to constitute a compact pool of classifiers
                end for
            end if
        end for
    end for
end for
```

The proposed ranking criterion is based on distance of the still ROI and the support vectors of e-SVMs dist $\left(\mathbf{ST}_{i,k,r}, \mathbf{sv}_{i,k,r}\right)$ in the feature space. It is assumed intuitively that those subspaces used for training are the most relevant ones, where the corresponding e-SVM classifiers have a larger distance to the target still than others. Subspaces are thereby ranked in descending order based on distance between the target still $\mathbf{ST}_{i,k,r}$ and e-SVM support vectors $\mathbf{sv}_{i,k,r}$ in the feature space (see Figure 4.3). N_{rs} set the number of over-produced subspaces, and N'_{rs} be the number of ranked subspaces.

4.1.4 Operational Phase (Dynamic Classifier Selection and Weighting)

An important challenge is to derive accurate measures for classifier competence in the context of the SSPP problem. The proposed approach allows the still-to-video FR system to select the classifiers that are most competent for the capture conditions. A new distance-based DS approach is proposed to provide the best classifiers to discriminate between the target and non-target ROIs. In order to dynamically select the most competent classifiers for the design of a robust ensemble, the proposed internal criteria (levels of competence) per given probe ROI relies on the: (1) distance from the non-target support vectors ROI patterns, and (2) closeness to the target still ROI pattern. The key idea is to select the classifiers that effectively locate the given probe ROI pattern close to the target still in the feature space. If the distance between the probe and the target still ROI pattern is lower than the distance to support vectors, then those classifiers are selected dynamically as competent classifiers for the given probe ROI pattern.

The distance from support vectors can be defined based on the distance to the closest support vector to the target still. On the other hand, the classifiers with support vectors that are far from the ROI test patterns of individuals of interest can be also suitable candidates, because they may classify probe ROI patterns correctly. In the proposed DS approach (illustrated in Figure 4.3), all the non-target support vectors were sorted based on their distance to the target still (the target support vector) in an offline processing. Then, the closest support vector to the target still is used to compare with the input probe.

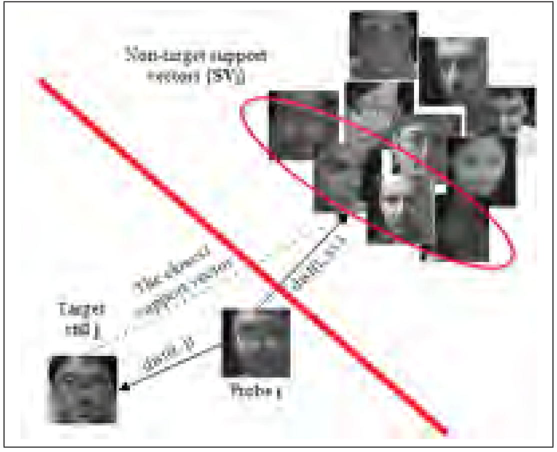


Figure 4.3 A 2-D illustration of the proposed dynamic classifier selection approach in the feature space.

During operations, each given probe ROI **t** is projected in the feature space and those classifiers form the pool that verify the selection criteria (locate the input near the target still and far from support vectors) are selected dynamically, and their scores are combined using score-level fusion. In contrast to the approaches that use local neighborhood accuracy for measuring the level of competence, it is not mandatory in the proposed method to define neighborhood using all the validation data, like with method based on, e.g., kNN. Thus, different distance metrics, such as Euclidean, CityBlock, Hamming, etc., can be employed to measure the distances between ROI patterns and support vectors. The algorithm of proposed classifier selection method is formalized in Algorithm 4.5.

Algorithm 4.5 Operational phase with DS.

```
Input: Pool of e-SVM classifiers $E_j$ for individual of interest $j$, the set of support vectors $\{SV_j\}$ per $E_j$
Output: Scores of dynamic ensembles based on a subset of the most competent classifiers $C_j^*$
for each probe ROI $t$ do
    Divide testing ROI $t$ into patches after preprocessing
    for each patch $i = 1...N_p$ do
        for each face descriptor $k = 1...N_{fd}$ do
            $\mathbf{a}_{i,k} \leftarrow$ extract features $f_k$ from patch $p_i$
            for each subspace $r = 1...Nrs$ do
                $\mathbf{a}_{i,k,r} \leftarrow$ sample subspaces $s_r$ from $RA_{s,j}$
                $C_j^* \leftarrow \{\emptyset\}$
                for each classifier $c_l$ in $C_j$ do
                    if dist$\left(\mathbf{a}_{i,k,r}, \mathbf{ST}_{i,k,r}\right) \leq$ dist$\left(\mathbf{a}_{i,k,r}, \mathbf{sv}_{i,k,r}\right)$ then
                        $C_j^* \leftarrow c_l \cup C_j^*$
                    end if
                end for
            end for
        end for
    end for
    if $C_j^*$ is empty then
        $S_j^* \leftarrow$ Use mean scores of $E_j$ to classify $t$
    else
        $S_j^* \leftarrow$ Use mean scores of $C_j^*$ to classify $t$
    end if
end for
```

As described in the Algorithm 4.5, each given input ROI t is first divided into patches p_i. Then, feature extraction technique f_k is applied on each patch to form a feature vector $\mathbf{a}_{i,k}$ per patch. Afterwards, the ranked subspaces stored in the $RA_{s,j}$ are sampled from $\mathbf{a}_{i,k}$ and then $\mathbf{a}_{i,k,r}$ is projected into the feature space containing support vectors $\{SV_j\}$ of classifiers and the reference still $\mathbf{ST}_{i,k,r}$ of target individual j. Finally, those classifiers c_l in E_j that satisfy the levels of competence criteria (line 13) are selected to constitute C_j^* in order to classify testing sample $\mathbf{t}$. Subsequently, the scores of selected classifiers $S_{i,k,r}$ are combined using mean function to provide final score S_j^*. All the classifiers in E_j are combined to classify ROI t when none of classifier fulfill the competence criteria.

In the proposed system, the ground-truth tracks are also exploited allowing to accomplish a robust spatio-temporal recognition. To that end, ROI captures for different individuals are regrouped through facial trajectories. In particular, decision fusion module accumulates the scores S_j^* of each individual-specific ensemble over a fixed size window W to make a decision d_j^* as follows:

$$d_j^* = \sum_{w=0}^{W-1} S_j^* \left[S_{i,k,r(W-w)} \right] \in [0, W] \tag{4.5}$$

Dynamic weighting of e-SVMs is suitable for rapid adaptation of individual-specific ensembles to tackle the variations within the operational domains. In this case, a distance-based combination strategy is also proposed to dynamically weight the scores of e-SVMs, where it relies on the distance of the probe instance to the support vectors of each classifier, as well as, to the target reference still in the feature space. This approach aims to reduce the effect of non-competent classifiers when their support vectors are closer to the given probe than the target still. Higher weights are assigned to the scores of classifiers with larger distance to the probe with respect to closeness to the single target still, and vice versa. Hence, each probe ROI pattern is compared to that of the single target still, and to that of the support vector of each classifier. If distance with the target still is closer than the closest support vector, then those classifiers are attributed higher weights. The proposed DW strategy is formalized in Algorithm 4.6.

Algorithm 4.6 Dynamic classifier weighting strategy.

```
1: Compute the distances of the probe with the closest support vector of each e-SVM and the target still,
   then store these distances dist$(t,sv)$, and dist$(t,j)$, respectively
2: Weight the scores of a classifiers $s_k$ and create the weighted scores $s_k^w = s_k.w_k$, where the $w_k$
   is the relative competence of the classifier $c_k$ on its corresponding weighted scores $s_k^w$ estimated as
   $w_k = \frac{\mathrm{dist}(t,sv)^2}{\mathrm{dist}(t,sv)^2+\mathrm{dist}(t,j)^2}$
3: Use the mean fusion of weighted scores $s_k^w$ to obtain the final score after score normalization
```

4.2 Experimental Results and Discussions

Several aspects of the proposed system are assessed experimentally using real-world video surveillance data. First, different e-SVM training schemes are compared for the individual-specific ensembles. Second, different pool generation scenarios are evaluated in terms of accuracy and time complexity. Finally, the impact of applying DS and DW are analyzed on the performance.

4.2.1 Experimental Protocol

In experiments on COX-S2V, the high-quality stills for $N_{wl} = 20$ individuals are randomly chosen to populate the watch-list, as well as, $N_{wl} = 10$ for evaluation of different training schemes. In addition, N_{ntd} video sequences of non-target persons from the OD are selected as calibration videos for the design phase. Moreover, N_{ntu} video sequences of unknown persons are considered for the operational phase. Hence, different subsets of COX-S2V are separated as demonstrated in Figure 4.4 according to design scenarios, validation, and operational phases of the proposed system. Validation set D as required in the first design scenario is separated to define the system parameters containing $N_a = 20$ stills and videos of some random individuals along with $N_{ntd} = 100$ (to calibrate for cameras and scores) and $N_{ntu} = 100$ testing videos of other unknown persons for the design and operational phases, respectively. Design set to create facial models (generating a pool of classifiers) including high-quality stills of watch-list individuals $N_{wl} = 20$ and low-quality calibration videos of non-target persons $N_{ntu} = 100$. Operational set (test set) to assess the system performance that consists of N_{ntu} videos belonging to another set of unknown persons, as well as, videos of a target individual. During operations, one target individual is considered at a time along with non-targets in the operational scene. In order to achieve statistically significant results, these experiments are replicated 5 times with considering different stills and videos of individuals of interest as watch-list persons.

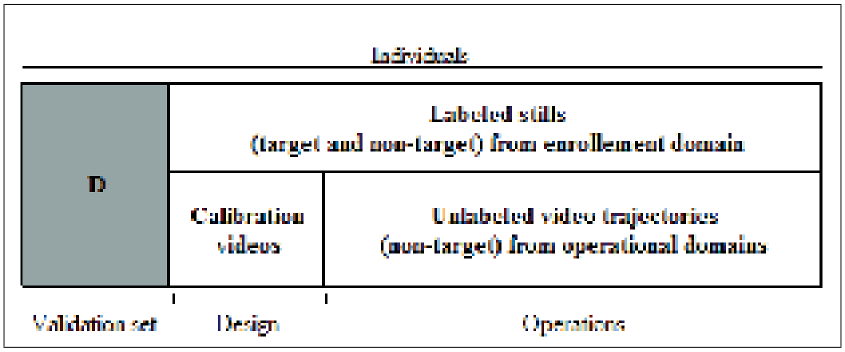

Individuals
D
Labeled stills
(target and non-target) from enrollement domain
Calibration
videos
Unlabeled video trajectories
(non-target) from operational domains
Validation set
Design
Operations

Figure 4.4 The separation of COX-S2V dataset for validation, design and operational phases of the proposed system.

In experiments on Chokepoint, stills of $N_{wl} = 5$ individuals of interest are considered to constitute the watch-list. Videos of $N_{ntd} = 10$ unknown persons are used as calibration videos to construct a pool of e-SVM classifiers, and videos of $N_{ntu} = 10$ other non-target individuals are associated for the operations along with videos of watch-list individuals.

The facial ROIs appearing in reference stills and video frames were isolated in the COX-S2V and Chokepoint using the viola-Jones face detection. The reference stills and video ROIs are all converted to grayscale and scaled to a common size of 48x48 pixels for computational efficiency (Huang *et al.*, 2015). Histogram equalization is used to enhance contrast, as well as, to eliminate the effect of illumination changes. Then, an uniform non-overlapping patch configurations is applied to divide each ROI into 9 blocks of 16x16 pixels as in (Bashbaghi *et al.*, 2015; Chen *et al.*, 2015). HOG and LPQ feature extraction techniques are utilized to extract discriminating features with the dimensions of 192 and 256, respectively. For HOG face descriptor, 3x3 pixel cells are considered with unsigned gradients, spacing stride of $l = 2$, and the default value of L2-Hys threshold. In addition, numbers and dimensions of feature subspaces are shown in Figure 4.5. Libsvm library (Chang & Lin, 2011) is used in order to train e-SVMs, where the same regularization parameters $C_1 = 1$ and $C_2 = 0.01$ are considered for all exemplars (**w** of a target sample is 100 times greater than non-targets) (Bashbaghi *et al.*, 2015). Random subspace sampling with replacement is also employed to generate different subspaces randomly from feature space.

Ensemble of template matchers (TMs) and e-SVMs using multiple face representations (Bashbaghi *et al.*, 2014, 2015), specialized kNN adapted for video surveillance (VSkNN) (Pagano *et al.*, 2014), sparse variation dictionary learning (SVDL) (Yang *et al.*, 2013), and ESRC-DA (Nourbakhsh *et al.*, 2016) are considered as the base-line and state-of-the-art FR systems to validate the proposed system. In kNN experiment, PCA is applied for ROIs (Zhang *et al.*, 1997) are employed to compute the VSkNN using $k = 3$ (1 target still from the cohort model along with 2 nearest non-target video ROIs). To that end, distances of the probe ROI t are calculated from the target still ST_j, as well as, two nearest non-target T_1 and T_2 from the calibration videos. Thus, VSkNN score (S_{VSkNN}) is obtained as follows (Pagano *et al.*, 2014):

$$S_{VSkNN} = \frac{\text{dist}(t, ST_j)}{\text{dist}(t, ST_j) + \text{dist}(t, T_1) + \text{dist}(t, T_2)} \tag{4.6}$$

where $\text{dist}(t, ST_j)$ is the distance of the probe face t from the target still ST_j, $\text{dist}(t, T_1)$ and $\text{dist}(t, T_2)$ are the distances of the given probe t from the two nearest non-target captures, respectively.

In SVDL experiment, high-quality stills belonging to the individuals of interest are considered as a gallery set and low-quality videos of non-target individuals are employed as a generic training set to learn a sparse variation dictionary. Three regularization parameters λ_1, λ_2, and λ_3 set to 0.001, 0.01, and 0.0001, respectively, and also the dimensionality of faces is reduced to 90 using PCA according to the default values defined in (Yang *et al.*, 2013). The number of dictionary atoms are initialized to 100 based on the number of stills in the gallery set, where it is a trade-off between the computational complexity and the level of sparsity.

4.2.2 Computational Complexity

In practical video surveillance applications, FR systems must be computationally efficient, and scale well to a growing number of cameras, watch-list individuals, and clutter in the scene. The generation of e-SVM classifiers comprised of training e-SVMs, ranking patches and subspaces, as well as, pruning the e-SVMs were performed off-line. Since e-SVMs trained for different patches, descriptors, and random subspaces are generated and ranked independently from one another, they can be processed in parallel. Computational complexity of the proposed system is therefore relevant to the operational phase, and affected by the feature extraction techniques, classification process, and dynamic selection and weighting of each input ROI probe with the size of nxn.

Extraction of face descriptors using HOG and LPQ is related to their transformation functions, where their complexities are $O(n)$ and $O(nlogn)$, respectively (Ahonen *et al.*, 2008; Deniz *et al.*, 2011). Classification has been performed using e-SVM which employs a linear SVM kernel function using a dot product with the complexity of $O(N_d \cdot N_{sv})$ (Chang & Lin, 2011), where N_d and N_{sv} are the average dimensionality of the face descriptors and the average number of support vectors, respectively. Finally, dynamic selection and weighting is based on City-block distance which is a linear distance metric, therefore, this process requires $O(N_d \cdot N_c \cdot N_{sv})$ computations, where N_c is the total number of classifiers in the pool.

Memory complexity of the proposed system mainly depends on the number of watch-list persons N_{wl} and size of the pool. Thus, complexity of the pool (number of classifiers N_c) for each individual of interest can be considered as $O\left(N_p \cdot N_{fd} \cdot N_{rs}\right)$, where N_p is the number of patches, N_{fd} and N_{rs} are the number of face descriptors and the average number of random subspaces, respectively. Hence, the overall memory complexity can be computed as $O\left(N_{wl} \cdot N_p \cdot N_{fd} \cdot N_{rs} \cdot N_d\right)$. More specifically, the worst case of computational complexity of the proposed individual-specific Ee-SVMs in the operational mode to process an input ROI pattern can be formulated as $N_p \cdot N_{fd} \cdot N_{rs} \cdot N_{sv} \cdot N_d$ according to the dot products required by each e-SVM classifier.

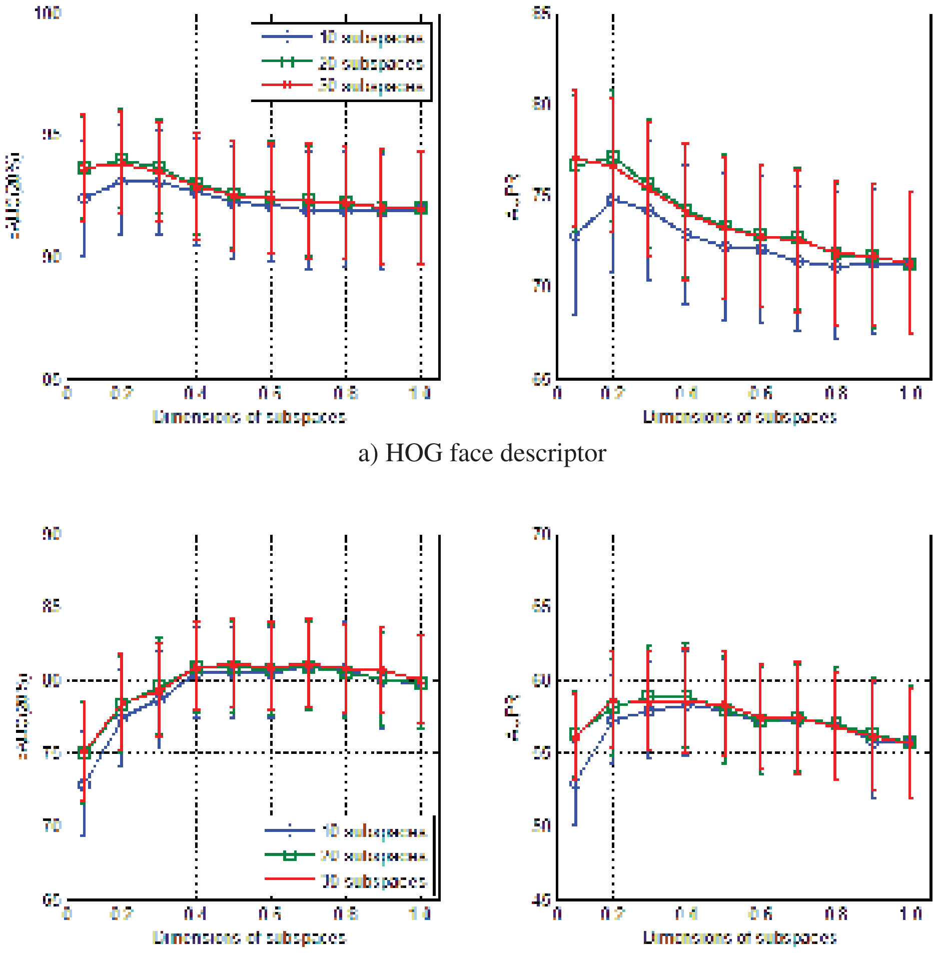
4.3 Results and Discussion

4.3.1 Number and Size of Feature Subspaces

The critical parameters of the proposed system need to be defined precisely to select the best values using the generic pool. The impact of different numbers and dimensions of feature subspaces are statistically analyzed for each face descriptors extracted from each patch using a validation set during the design phase. In this analysis, different numbers of subspaces (N_{rs}) are considered w.r.t. different proportions of feature dimensions (N_d). In this section, experiments were conducted with a generic pool that uses RSM to generate individual-specific Ee-SVMs combined through score averaging based on the third training scheme. The transaction-level analysis (pAUC(20%) and AUPR with standard errors) of different numbers and dimensions of subspaces for HOG and LPQ are depicted in Figure 4.5.

Figure 4.5 (a) implies that performance obtained using 20% of features is slightly higher than other dimensions in term of both pAUC(20%) and AUPR for HOG descriptor. Results suggest that it is better to select the 20% of original feature space as a dimensions of HOG descriptor (39 features). In addition, 20 random subspaces as the number of subspaces achieves the highest performance.

As shown in Figure 4.5 (b), 40% of the LPQ descriptor can be a suitable value as dimension of LPQ subspaces. Moreover, the best number of subspaces can be defined as 20 subspaces. It can be seen that performance of the system is not greatly affected by the numbers and dimensions of feature subspaces, where either pAUC(20%) or AUPR first raise and then stabilize. This suggests that increasing the number of subspaces may transfer more diversity among classifiers in the pool, but it cannot improve the accuracy. Noted that, performance is stabilized for the values higher than 20 subspaces. Hence, it can be concluded that the proposed system is not highly sensitive to the number of subspaces (see Figure 4.5 (a)).



a) HOG face descriptor

b) LPQ face descriptor

Figure 4.5 The impact of different numbers and size of feature subspaces on performance of using HOG and LPQ face descriptor.

Another experiment that was performed prior to design is to rank patches using the validation set *D*. The sensitivity analysis on the performance of using each patch separately in order to rank them based on their importance is illustrated in Figure 4.6.

As shown in Figure 4.6, each patch performs differently from other patches for each descriptor. Selecting a different number of semi-random subspaces from each patch based on its importance for overall performance therefore can lead to a robust system.

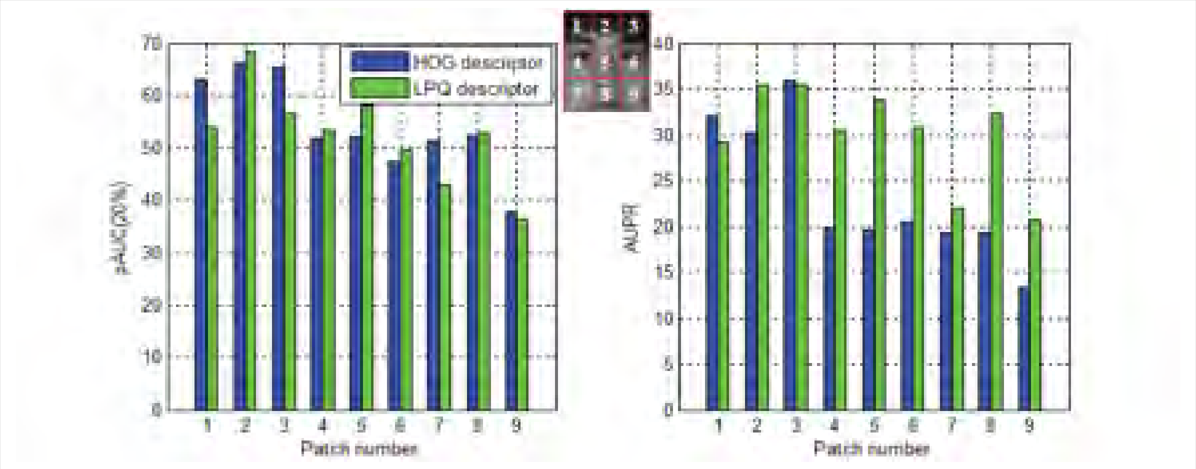


Figure 4.6 The analysis of system performance based on each patch over COX-S2V.

4.3.2 Training Schemes

Figure 4.7 presents the average transaction-level performance of using the generic pool for different training schemes as described in Section 4.1.2.2 over each video of COX-S2V. Results were produced using a generic pool of 360 e-SVMs (9 patches x 2 descriptors x 20 subspaces) per each target individual.

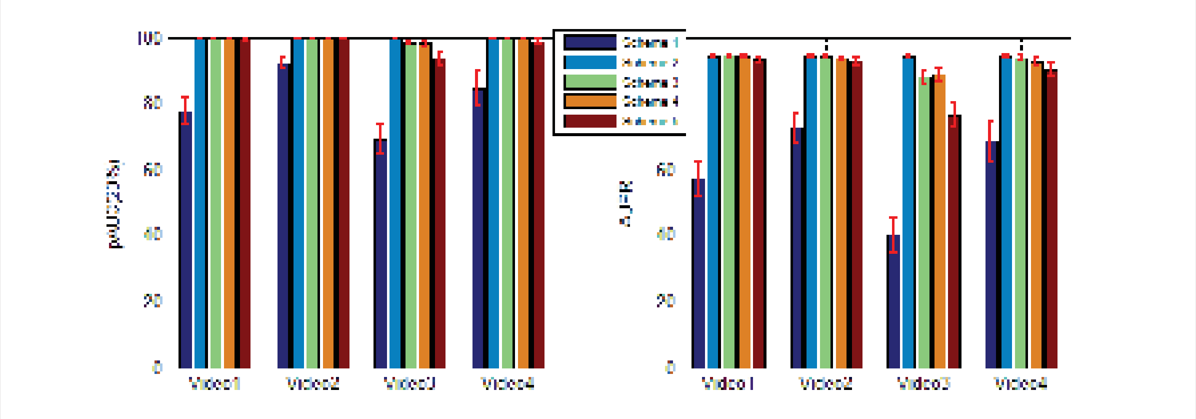


Figure 4.7 Average pAUC(20%) and AUPR transaction-level performance of different training schemes at with COX-S2V.

Results in Figure 4.7 indicate that training schemes 2, 3, 4, and 5 greatly outperform scheme 1, due to DA using knowledge transfered from all of the surveillance cameras in the target domain. The results also suggest that exploiting a few non-target stills from the source domain during training e-SVMs in the third scheme can provide slight improvements, especially in AUPR values according to video1, video2, and video4 comparing to the second scheme (Bashbaghi *et al.*, 2015). Knowledge of the ED is therefore incorporated in the third scheme due to combination of feature representations across domains using a mixture of labeled still ROIs from the ED and unlabeled calibration videos from the OD (Pan & Yang, 2010).

Camera-specific training schemes 4 and 5 provide higher performance in comparison to scheme 1, where they also exploit knowledge of the operational domain. However, they are also outperformed by schemes 2 and 3 in terms of both accuracy and complexity, because videos from all of the cameras are considered in schemes 2 and 3 to generate a general pool, while several camera-specific pools must be generated in the schemes 4 and 5 using videos of each specific camera. Meanwhile, scheme 4 performs slightly better than scheme 5, because all of the video ROIs captured from a specific camera FoV have the same pose and angle, while adding frontal stills with significant differences in data distributions may subsequently degrade the training performance. Noted that only the classifiers from pool #1 trained using camera #1 is employed to classify the probe captured using camera #1 during operational phase.

Therefore, other experiments on the proposed system are accomplished using the third training scheme. Since the characteristics of capturing devices are different, it has a significant impact on the system performance according to each video. The differences between pAUC(20%) and AUPR observed in Figure 4.7 reveal that the large number of e-SVMs classify the non-target ROIs as non-targets, but only some of them classify the target ROIs correctly. Therefore, the FPR values are very low in the all cases.

Another test that can be also used in order to assess the performance of the training schemes is the Friedman test with a post-hoc test, where it is basically incorporated to find a significance difference between several methods according to their ranks averaged across datasets. The

Friedman test is typically followed by a post-hoc test, such as Nemenyi test to indicate whether the difference in ranks is above a critical distance (CD) (Demsar, 2006). Figure 4.8 shows the results of Nemenyi's post-hoc test, where the schemes linked by colored lines are not significantly different by the test for a significance level of $\rho = 0067$.

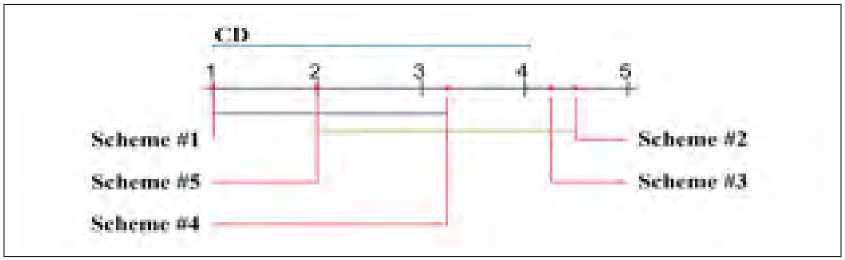


Figure 4.8 Training schemes by significant differences according to the post-hoc Nemenyi test over COX-S2V.

Figure 4.8 demonstrates with a more visual insight as differences of the training schemes, where the lowest average rank is associated to the worst training scheme and vice versa. According to this test, schemes that exploit DA are significantly different than scheme #1, meaning that training through DA provides significantly higher performance than the training without considering DA.

4.3.3 Number of Training and Testing Stills and Trajectories

The impact of employing different number of non-target videos from the background model (videos of non-target persons), as well as, different number of non-target stills from the cohort model (stills of non-target persons) on the performance is illustrated in Figure 4.9. In this regard, the third training scheme is employed considering the first $N_{wl} = 10$ persons of COX-S2V as watch-list individuals. The number of low-quality videos of non-target persons N_{ntd} considered for training during the design phase is varied from 10 to 100 according to the number of non-target stills belonging to other persons in the cohort.

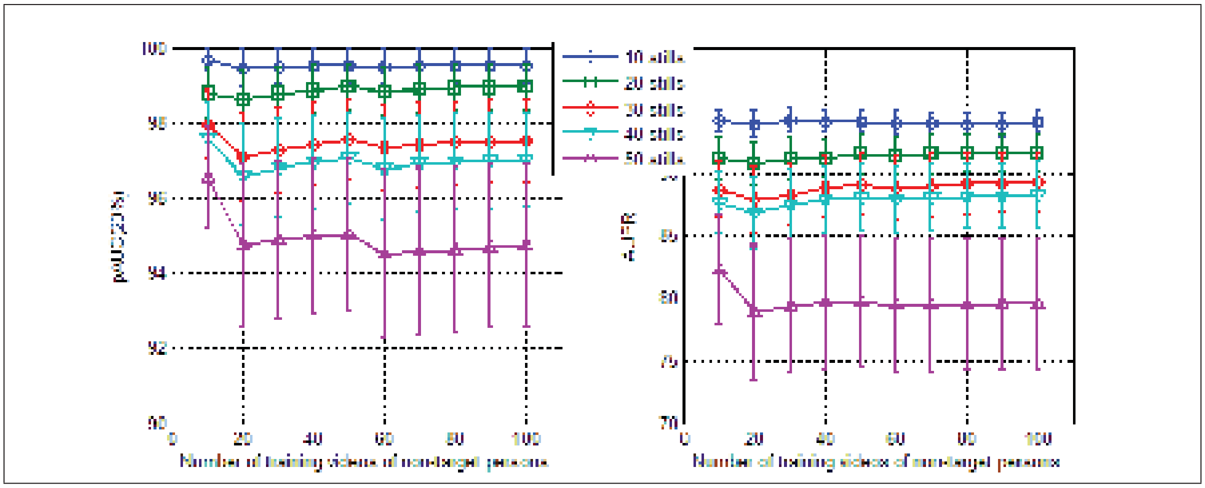


Figure 4.9 The analysis of system performance using different number of training non-target persons over COX-S2V.

As shown in Figure 4.9, growing the number of non-target persons participating in the design phase can slightly improve the performance. Since it may be costly and impractical to employ plenty of training data in the real-world application, the proposed system provides convincing results even with limited non-target video data. Thus, knowledge of the target domain can be appropriately transferred by considering the limited number of non-target video data.

Figure 4.9 also demonstrates that growing the number of high-quality non-target stills during training degrades the performance significantly. Since these still ROIs are close to the still of the target individual, most of the support vectors are selected from them and subsequently, these classifiers could not successfully classify the low-quality input probes. Hence, the larger the number of non-target stills, the higher the number of inappropriate support vectors, and therefore the capability of classifiers reduces to classify the given probe as if employing a lower number of non-target stills. Nevertheless, employing the lower number of stills from the cohort along with videos of non-target persons provides higher classification performance as shown in Figure 4.7.

To analyze the performance considering different number of watch-list individuals enrolled to the system, N_{wl} is varied from 5 to 20 as illustrated in 4.10.

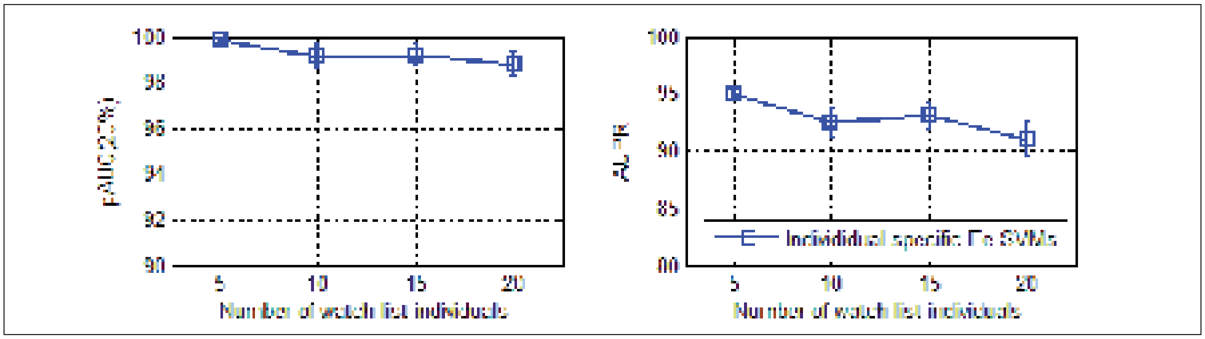


Figure 4.10 The analysis of Ee-SVMs performance using different number of watch-list persons during operations over COX-S2V.

Figure 4.10 shows that enlarging the list of watch-list persons does not have a significant impact on the system performance. Since the proposed system is comprised of individual-specific ensembles, and each one seeks to detect one watch-list individual at a time, there should not be significant differences in increasing the number of watch-list persons.

The impact of considering different number of non-target videos of unknown persons from the test set on performance is displayed in Figure 4.11. In this regard, the number of unknown persons N_{ntu} appearing in the surveillance environment along with the target person during the operational phase is altered to see its influence on the system performance.

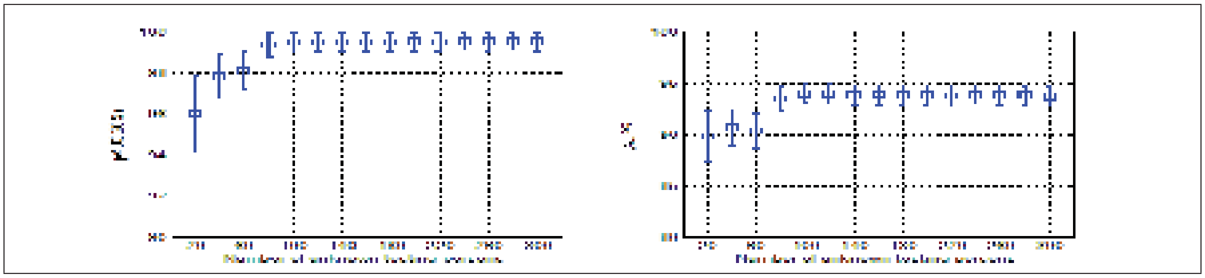


Figure 4.11 The analysis of system performance using different number of unknown persons during operations over COX-S2V.

As illustrated in Figure 4.11, the number of unknown persons participating in the operational phase is varied from 20 to 300 persons. Since the FP values for each threshold in the ROC and inverted precision-recall curves increase slower than the total number of negatives, then the FPR values decrease slightly and it subsequently leads to a higher values of area under ROC and precision-recall curves. It can be concluded that the proposed system can perform well even with severely imbalanced data according to observation of many unknown persons during operations.

To obtain the transaction-level performance of the proposed system using pAUC, values of FPR are varied from 5% to 100% as demonstrated in Figure 4.12.

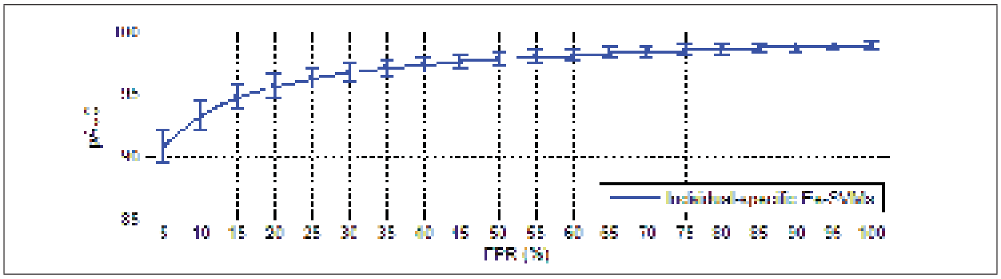


Figure 4.12 The analysis of transaction-level performance according to pAUC using different values of FPR over COX-S2V.

As shown in Figure 4.12, increasing the FPR thresholds can slightly achieve higher AUC, while the real-world watch-list screening systems must perform on a certain operating point that has been considered as FPR=20% in this chapter. Thus, the rate of false positives must be limited by considering an appropriate operating point w.r.t. the application.

4.3.4 Design Scenarios

Performance of the proposed system in terms of considering different design scenarios is presented in Table 4.1 and 4.2 using the third training scheme over videos of COX-S2V and Chokepoint, respectively.

The results in Table 4.1 indicate that generating a compact pool of classifiers based on the first design scenario can yield higher performance, where the baseline performance is obtained

Table 4.1 Average pAUC(20%) and AUPR performance of the system with generic pool and different design scenarios at transaction-level over COX-S2V.

Systems	**Video 1**		**Video 2**		**Video 3**		**Video 4**		**Complexity**
	pAUC	**AUPR**	**pAUC**	**AUPR**	**pAUC**	**AUPR**	**pAUC**	**AUPR**	**(# dot products)**
Generic pool	99.19±0.44	93.18±0.88	99.43±0.16	91.39±0.82	92.01±1.11	70.95±2.20	96.08±1.09	84.89±2.08	460,080
Scenario 1	99.97±0.03	94.86±0.18	99.40±0.22	92.60±0.78	97.77±0.52	87.23±1.13	93.12±0.90	81.18±0.87	127,800
Scenario 2	99.08±0.40	92.64±0.69	99.32±0.17	90.44±1.01	91.02±1.28	68.54±2.36	96.21±1.02	84.37±2.11	230,040

using the generic pool. Hence, pre-selection of e-SVMs by ranking patches and subspaces achieves better performance with a lower computational complexity. Moreover, system with a compact pool generated according to the second design scenario cannot improve the performance effectively, since no priori knowledge is taken into account and all system choices are performed during the enrollment phase. Consequently, generating a compact pool according to the first design scenario using the criteria based on overall AUPR through a validation set is more accurate and efficient.

Table 4.2 Average pAUC(20%) and AUPR performance of the system with generic pool and different design scenarios at transaction-level over Chokepoint.

Systems	**Session 1**		**Session 2**		**Session 3**		**Session 4**		**Complexity**
	pAUC	**AUPR**	**pAUC**	**AUPR**	**pAUC**	**AUPR**	**pAUC**	**AUPR**	**(# dot products)**
Generic pool	97.67±0.92	96.63±1.21	96.93±1.43	95.33±2.21	100±0.00	99.64±0.07	75.33±6.04	71.85±6.66	460,080
Scenario 1	99.74±0.12	99.25±0.25	99.99±0.01	99.81±0.01	100±0.00	99.74±0.05	91.81±0.92	90.56±1.16	127,800
Scenario 2	98.81±0.49	98.15±0.56	98.07±0.82	96.89±1.36	100±0.00	99.74±0.08	77.00±5.59	73.52±6.37	230,040

Table 4.2 also confirms that the results obtained with the first design scenario are higher than generic pool and compact pool generated according to the second design scenario among all the sessions. On the other hand, performance of the system with the second design scenario is slightly better than the baseline using the generic pool.

It is worth pointing out that, the number of classifiers in the generic pool for each individual of interest is 360 (9·2·20), while each target individual has about 100 and 180 classifiers in the compact pool of first and second scenarios. Meanwhile, the average number of support vectors for each classifier and the dimension of each feature vector are 18 and 71, respectively. Thus, the time complexity as described in Section 4.2.2 for generic pool is about 460,080 (360·18·71) dot products for processing a given probe ROI, while the compact pool based on the first and second scenarios requires around 127,800 and 230,040 computations, respectively. Hence, the proposed system based on the first scenario is effective in terms of either accuracy or computational complexity.

Furthermore, the impact of different numbers of ranked subspaces in the system with a pool generated based on the second design scenario is shown in Figure 4.13. In this scenario, over-produce and select paradigm is considered, where 50 subspaces are generated for each patch and then they are ranked using the local distance-based criteria (see Section 4.1.3).

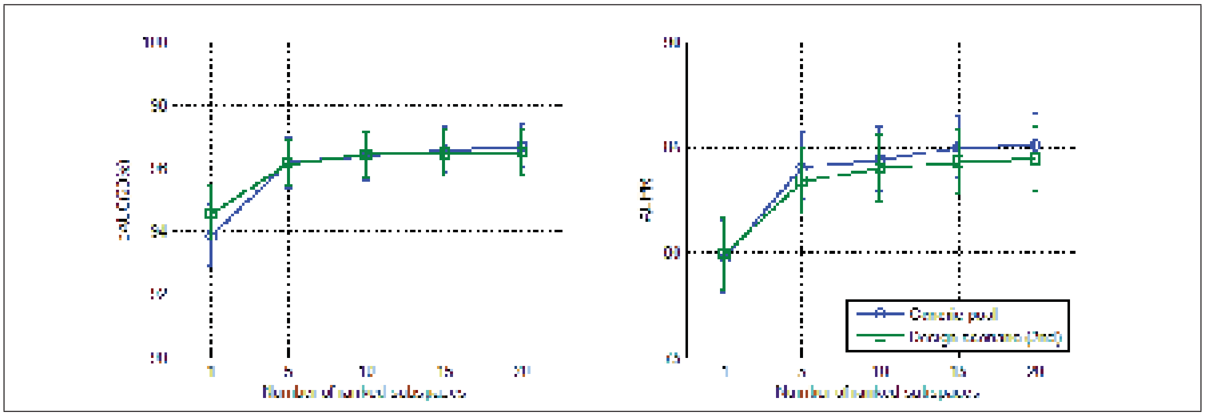


Figure 4.13 The analysis of system performance using different numbers of ranked subspaces based on the second design scenario of compact pool generation over COX-S2V.

As shown in Figure 4.13, both systems perform equally in terms of pAUC(20%) values, while the system designed with the second scenario outperforms the generic pool specifically for the first 10 ranked subspaces. Moreover, the pAUC performance is stable starting from $N'_{rs} = 10$ subspaces. The system with the generic pool performs better in terms of AUPR values. It can be concluded that the local criteria exploited to select the best subspaces in the second design scenario cannot be a desired metric consistently in contrast to the global criteria utilized in the first design scenario.

The diversity among classifiers within each individual-specific ensemble is computed using kappa (k) diversity measure for $N_{wl} = 20$ individuals with 5 replications. The value of k is

0.0065±0.0005, where it can be concluded that the classifiers within the ensembles are relatively diverse.

4.3.5 Dynamic Selection and Weighting

The performance of applying dynamic selection and weighting approaches on the proposed system with generic pool, the first design scenario (compact pool), and the second design scenario are demonstrated in Table 4.3 using the third training scheme at transaction- and trajectory-level along with the time complexity.

Table 4.3 Average pAUC(20%) and AUPR performance at transaction- and trajectory-level after applying dynamic selection and weighting on the system with generic pool and different design scenarios over COX-S2V.

Systems	**Transaction-level**		**Trajectory-level**	**Complexity**
	pAUC	**AUPR**	**AUC**	**(# dot products)**
Generic pool	96.68±0.70	85.10±1.49	99.72±0.05	(9·2·20·18·71) = 460,080
Generic pool with DS	98.21±0.45	86.40±1.17	99.93±0.04	(9·2·20·18·71) + (9·2·20·2·71) = 511,200
Generic pool with DW	97.52±0.59	87.27±1.38	99.91±0.04	(9·2·20·18·71) + (9·2·20·2·71) = 511,200
Generic pool with DS and DW	96.89±0.64	85.39±1.47	99.90±0.05	(9·2·20·18·71) + 2·(9·2·20·2·71) = 562,320
Scenario 1 with DS	93.47±0.76	77.32±1.66	99.52±0.14	(100·18·71) + (100·2·71) = 142,000
Scenario 1 with DW	98.11±0.49	88.60±1.24	99.93±0.05	(100·18·71) + (100·2·71) = 142,000
Scenario 1 with DS and DW	95.60±0.72	84.08±1.39	99.77±0.10	(100·18·71) + 2·(100·2·71) = 156,200
Scenario 2 with DS	98.02±0.47	86.14±1.25	99.87±0.07	(9·2·10·18·71) + (9·2·10·2·71) = 255,600
Scenario 2 with DW	97.38±0.82	87.36±1.84	99.89±0.05	(9·2·10·18·71) + (9·2·10·2·71) = 255,600
Scenario 2 with DS and DW	96.37±0.98	85.12±1.88	99.76±0.08	(9·2·10·18·71) + 2·(9·2·10·2·71) = 281,160

Table 4.3 indicates that applying proposed DS method can improve the performance in contrast to combining all of classifiers in the system with generic pool and the second design scenario. It implies that combining a subset of competent classifiers leads to a system with higher accuracy. In addition, the proposed DW approach performs better in comparison with DS in terms of AUPR, where only two distances (distance to the target still and distance to the closest non-target support vector) are measured in the both selection and weighting strategies. Moreover, applying the proposed DW approach on the scores of classifiers selected dynamically cannot achieve a better performance, due to elimination of classifiers.

As observed in Table 4.3, DW can also magnify performance of the system with the first design scenario slightly, while applying the dynamic selection approach deteriorates its performance. Since a pre-selection scenario was already applied to the compact pool, applying DS can diminish the ensemble diversity. It can be concluded that using the compact pool and weighting the classifiers dynamically achieves the highest level of performance considering the AUPR values.

The trajectory-level performance of the proposed systems with DS and DW are also presented in Table 4.3 as the result of spatio-temporal FR. Thus, scores of individual-specific ensembles are gradually accumulated over a window of $W = 10$ consecutive frames using a trajectory defined by the tracker. To assess the overall performance, the corresponding ROC curve can be then plotted for each individual of interest by varying the thresholds from 0 to 10 (size of the window) over the accumulated scores, and the AUC are computed as overall performance.

As shown in Table 4.3, spatio-temporal recognition applied on the proposed systems leads to a near perfect face screening system. An example of accumulated scores over the generated trajectory is shown in Figure 4.14 using the systems with the best AUPR values. Video1 of COX-S2V is thus employed in this example, where individual ID#001 is considered as the watch-list target individual along with $N_{ntu} = 100$ unknown non-target individuals.

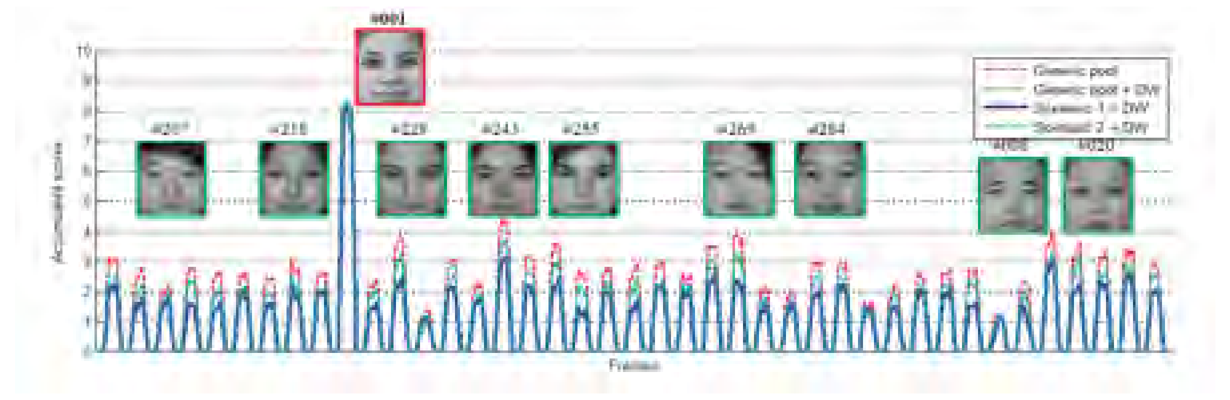


Figure 4.14 An example of the scores accumulated over windows of 10 frames over video1 of COX-S2V.

As shown in Figure 4.14, the accumulated scores for target individual (ID#001) is significantly higher than all non-targets individuals. It can be observed that the accumulated scores of some non-target individuals are high, due to appearance similarity to the target individual. The proposed system based on the first scenario with DW performs more reliable in trajectory-level, where it provides higher accumulated scores for the target, and simultaneously lower accumulated scores for non-target individuals.

Table 4.4 presents the complexity in terms of the number of dot products required during operations to process a probe ROI. The proposed selection and weighting approaches are desirable for the screening application in terms of operational time complexity. On the other hand, the distance measures can influence on the computational time based on their complexity. However, the CityBlock distance measure can be a suitable candidate due to its efficiency and linear computability. For example, the proposed system with DW over COX-S2V data needs 9·2·20·18·71 dot products for fusion in the worst case, where all of the classifiers are dynamically selected, and 9·2·20·2·71 for selection. It is worth pointing out that the average number of support vectors N_{sv} (the fourth element in the complexity formulation) for COX-S2V and Chokepoint data are not the same (18 and 14 support vectors, respectively), so that the computational complexity over these datasets is different.

Results are compared with the state-of-the-art and baseline systems in Table 4.4 according to the average transaction-level performance over the COX-S2V and Chokepoint data.

Table 4.4 Average pAUC(20%) and AUPR performance and time complexity of the proposed system at transaction-level over COX-S2V and Chokepoint videos against the state-of-the-art systems.

Systems	**COX-S2V**			**Chokepoint**		
	pAUC	**AUPR**	**Complexity**	**pAUC**	**AUPR**	**Complexity**
VSkNN (Pagano *et al.*, 2014)	56.80±4.02	26.68±3.58	671,744	19.00±0.40	16.48±0.90	671.744
SVDL (Yang *et al.*, 2013)	69.93±5.67	44.09±6.29	810,000	74.91±4.03	65.09±4.82	810,000
ESRC-DA (Nourbakhsh *et al.*, 2016)	99.00±1.13	63.21±4.56	228,614,400	97.16±1.28	76.97±6.73	432,224,100
Ensemble of TMs (Bashbaghi *et al.*, 2014)	84.00±0.86	73.36±9.82	1,387,200	85.60±1.04	82.78±7.06	1,387,200
Ensemble of e-SVMs (Bashbaghi *et al.*, 2015)	99.02±0.15	88.03±0.85	2,281,472	100±0.00	99.24±0.38	2,235,392
Scenario 1 with DW	98.11±0.49	88.60±1.24	142,200	97.52±0.50	96.86±0.72	113,600
Scenario 2 with DW	97.38±0.82	87.36±1.84	255,600	93.36±1.97	91.79±2.45	204,480

It can be seen from Table 4.4 that ensemble of e-SVMs significantly outperforms ensemble of TMs, VSkNN, SVDL, and ESRC-DA. Performance of the screening system using VSkNN and SVDL is poor, mostly because of the notable differences between quality and appearances of the target face stills in the gallery set and video faces in the generic training set, as well as, severely data imbalance of target versus non-target individuals observed during operations. It is worth noting that both VSkNN and SVDL are more suitable for close-set FR problems, such as face identification. Since each face captured should be assigned to one of the target still in the gallery, therefore, many false positive occur. Moreover, SVDL can only apply as a complex global N-class classifier in contrast to the proposed ensemble of SVMs, due to sparse optimization and classification during the operational phase. However, sparsity concentration index (Wagner *et al.*, 2012) is used as a rejection threshold to reject the probes not appearing in the training.

The results observed from Table 4.4 confirm that the proposed system using the first design scenario along with DW approach is efficient and can achieve an equivalent performance comparing to (Bashbaghi *et al.*, 2015) with a significant decrease in computational complexity. In addition, the system design with the second scenario and DW can perform almost equivalent to state-of-the-art systems performance. However, the systems proposed in this chapter employ two different face descriptors, whereas ensemble of e-SVMs utilizes four different face descriptors along with PCA with $O\left(N_d^3\right)$ for feature selection. Meanwhile, ensemble of TMs and VSkNN employ Euclidean distance with $O\left(N_d^2\right)$ to calculate the similarity among templates, therefore, they need more computations.

The proposed system is also validated using Chokepoint dataset, where the results observed from Table 4.4 confirm that the proposed system can achieve promising performance compare to state-of-the-art systems with a significantly lower computational complexity.

A Friedman test is also conducted on the comparison of the proposed systems against state-of-the-art and found significant with a significance level of ρ-value $\rho = 0.012$. The results of the Nemenyi post-hoc test is shown in Figure 4.15. These systems are ranked in an ascendant order, where the highest average rank is assigned to the best system. It indicates that the other four systems (ranked 1 to 5) are not significantly different, while the proposed system using design scenario 1 with DW is slightly different than the others and above the critical distance.

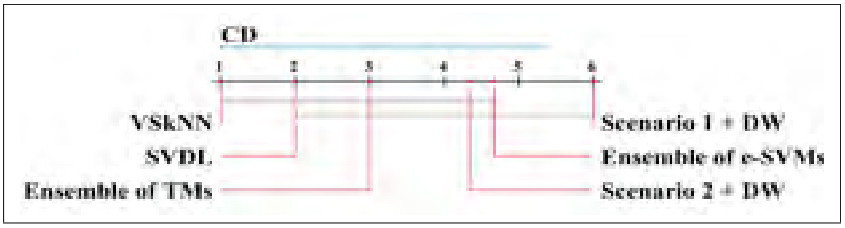


Figure 4.15 State-of-the-art systems by rank and differences according to the post-hoc Nemenyi test over COX-S2V.

CHAPTER 5

DEEP FACE-FLOW AUTOENCODERS FOR STILL-TO-VIDEO FACE RECOGNITION FROM A SINGLE SAMPLE PER PERSON

To improve the performance of video-based FR systems trained with a SSPP, deep neural networks have been recently proposed for extracting robust and nonlinear feature representations (Yang *et al.*, 2017; Hong *et al.*, 2017; Gao *et al.*, 2015). For example, robust convolutional features have been extracted in (Yang *et al.*, 2017) by sampling and detecting facial points using CNNs integrated with a joint and collaborative sparse representation based classification (SRC). A deep DA network with generating synthetic pose-free faces using a 3D face model has been introduced in (Hong *et al.*, 2017) to tackle the SSPP constraints. Moreover, autoencoder neural networks have shown to extract deterministic nonlinear feature mappings robust to face images contaminated by different noises, such as illumination, expression and poses (Gao *et al.*, 2015; Parchami *et al.*, 2017c).

In the CNN-based FR literature, it has been shown that deep recognition models trained on only still images or videos cannot perform convincingly on the other (Bansal *et al.*, 2017). In spite of their great success in FR with SSPP, they are not desirable for still-to-video FR because of their computational complexity and also discrepancies in the domains of still and video images. Subsequently, an accurate deep model requires to simultaneously consider both still images and videos during training and optimization of the network. Nevertheless, a canonical face representation through a supervised autoencoder (CFR-CNN) has been proposed by the authors in (Parchami *et al.*, 2017c) as the baseline still-to-video FR system that considers DA to reconstruct frontal faces from video ROIs. A fully-connected network was separately trained to classify the input probes.

Motivated by the effectiveness of autoencoders to remove the facial variations, a supervised end-to-end autoencoder is proposed in this chapter that considers both still images and videos to train of the network. In particular, the Face-Flow autoencoder CNN (FFA-CNN) is developed to deal with the SSPP problem in still-to-video FR, as well as, to restrain the differences between the enrollment and operational domains in the context of DA. The proposed FFA-CNN is trained using a novel weighted loss function to incorporate reconstruction and classification networks in order to recover high-quality frontal faces without blurriness from the low-quality video ROIs, and also to generate robust still and video representations similar for the same individuals through preserving identities to enhance matching capabilities. Therefore, the perturbation factors encountered in video surveillance environments and also the intra-class variations commonly exist in the SSPP scenario can be tackled using supervised end-to-end training.

The main contributions of this chapter lie in threefold: Firstly, a new type of supervised autoencoder network is adapted for single sample still-to-video FR that can be trained end-to-end with the reconstruction and classification networks. Secondly, video ROIs taken from the target domain are used simultaneously along with still ROIs captured from the source domain during training to address the problem of DA. Finally, a combinatorial weighted loss function including pixel-wise, symmetry-wise and identity preserving losses is exploited in the training process to learn robust still and video representations that can handle different variations, such as intra-class variances, self-occlusion and large poses. The proposed FFA-CNN is evaluated over the challenging Cox Face DB (Huang *et al.*, 2015), where it can yield comparable results to the baseline and state-of-the-art FR systems with lower number of training data.

5.1 Face-Flow Autoencoder CNN

The proposed FFA-CNN is an deep learning architecture containing a frontal face reconstruction network and a classification network. The reconstruction network (see Figure 5.1) transforms a pair of consecutive low-quality video ROIs to a well-illuminated frontal face with neutral-expression. It is designed using a supervised autoencoder network to eliminate the variations existing in video ROIs and yield a canonical noise-free face images. In contrast, the fully-connected classification network compares the face representations of a pair of consecutive video ROIs and a single still ROI to obtain a matching score. The reconstruction and classification networks are trained through supervised end-to-end multi-task learning to improve the performance of both face reconstruction and matching tasks. Therefore, the proposed FFA-CNN makes use of still and video ROIs simultaneously through DA and provides discriminant representations for robust still-to-video FR.

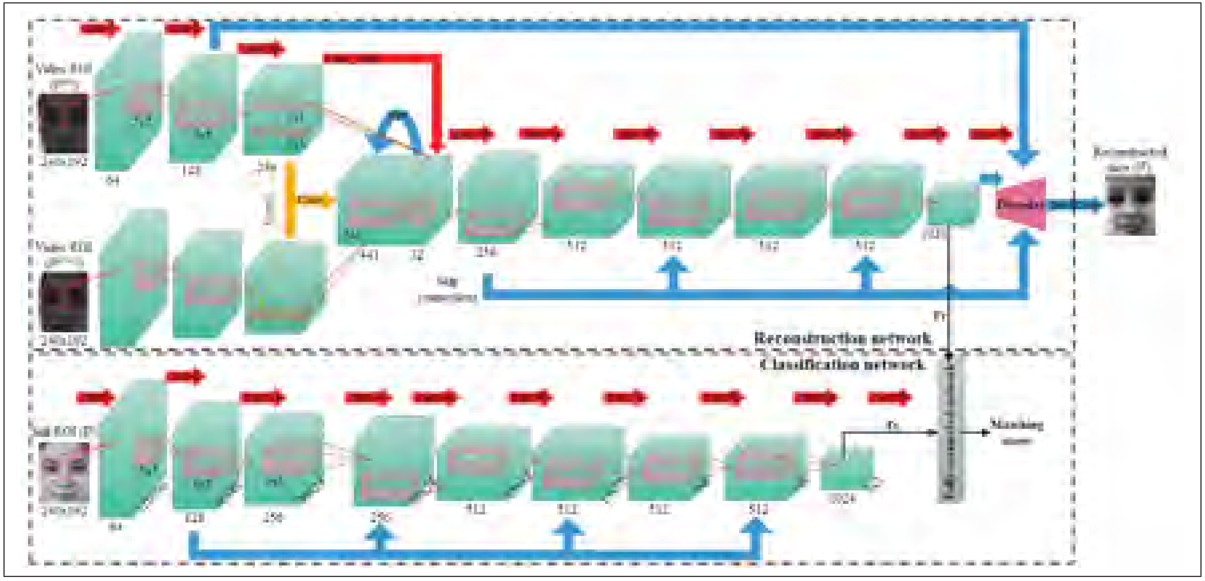


Figure 5.1   The reconstruction and classification networks of the proposed FFA-CNN.

In the proposed architecture that is inspired by FlowNet (Dosovitskiy *et al.*, 2015; Mayer *et al.*, 2016), a correlation layer is exploited in an end-to-end supervised fashion to enhance the capabilities of network to match different pairs of consecutive video ROIs. Thus, these matches are predicted at multiple levels to learn the variations among consecutive video ROIs and simultaneously generate a higher resolution face to be matched with the high-quality frontal still ROI. It allows to learn discriminative features at several scales, and subsequently a robust correspondence among pairs of video ROIs and the still ROI using different loss functions. It ensures that the network can perform effectively on realistic video-based FR scenarios without any additional face alignment and fine-tuning. In addition, this can be performed for each camera FoV during a calibration process.

5.1.1 Reconstruction Network

The aim of reconstruction network is to reconstruct a frontal face I^F using a pair of consecutive video ROIs $I^{p(t)}$ and $I^{p(t+1)}$. As shown in Figure 5.1, this pair of video ROIs are fused through a correlation layer to generate a video representation $r_v = E_v\left(I^{p(t)}, I^{p(t+1)}\right)$. The correlation layer is considered to effectively combine the feature maps extracted from $I^{p(t)}$ and $I^{p(t+1)}$ through multiplication.

Upconvolutional layers are employed in the decoder (see Figure 5.2) to perform upsamling of r_v due to reconstruction of a higher resolution frontal face image I^F. Skip connections are also considered for propagating the high-resolution information among the subsequent layers to preserve detailed information of faces (He *et al.*, 2016). To that end, the feature maps in the decoder are up-convolved and the outputs of upconvolution are concatenated with corresponding feature maps obtained from the encoder part to transfer high-level information to the reconstruction process.

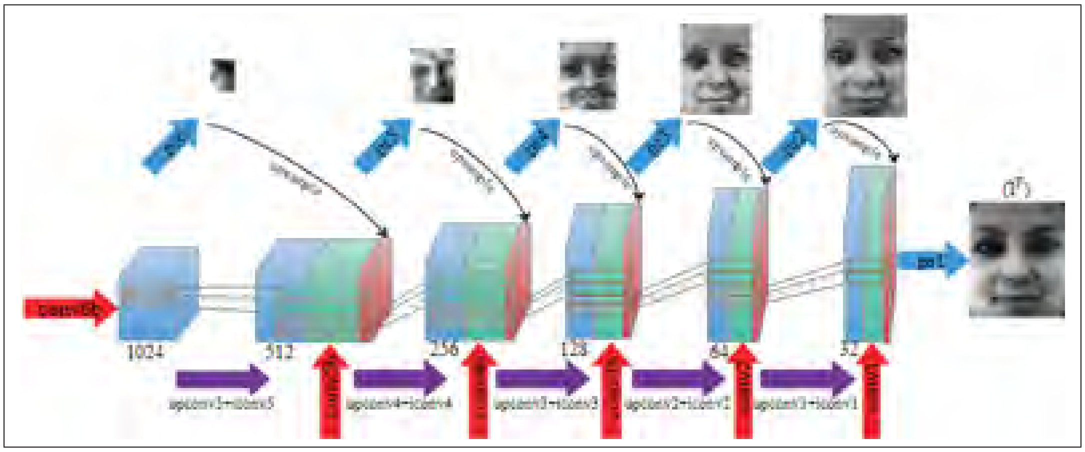


Figure 5.2 The decoder of the proposed FFA-CNN.

The architecture of reconstruction network is shown in Figure 5.1. It consists of ten strided convolutional and ten deconvolutional layers, as well as, one correlation layer. Note that conv1, conv2 and conv3 layers are applied to both $I^{p(t)}$ and $I^{p(t+1)}$. Specifications of the proposed network are presented in Table 5.1. As presented in Table 5.1, the network provides six predictions ($pr1 \ldots pr6$) at different scales that are used in the loss function to preserve reconstruction details. Identity convolutions ($iconv1 \ldots iconv5$) are regular convolutional layers with stride of 1 applied to the corresponding concatenated inputs.

Table 5.1 Specifications of the proposed framework.

Layer	KerSize/Str	Ch I/O	InputRes	OutputRes	Input
conv1	7x7/2	1/64	240x192	120x96	faces
conv2	5x5/2	64/128	120x96	60x48	conv1
conv3a	5x5/2	128/256	60x48	30x24	conv2
conv_redir	1x1/1	256/32	30x24	30x24	conv3a
corr	3x3/1	256/441	30x24	30x24	conv3a, conv3a[1]
conv3b	3x3/1	473/256	30x24	30x24	corr+conv_redir[2]
conv4a	3x3/2	256/512	30x24	15x12	conv3b
conv4b	3x3/1	512/512	15x12	15x12	conv4a
conv5a	3x3/2	512/512	15x12	7x6	conv4b
conv5b	3x3/1	512/512	7x6	7x6	conv5a
conv6a	3x3/2	512/1024	7x6	3x3	conv5b
conv6b	3x3/1	1024/1024	3x3	3x3	conv6a
pr6	3x3/1	1024/1	3x3	7x6	conv6b
upconv5	4x4/2	1024/512	7x6	15x12	conv6b
iconv5	3x3/1	1024/512	15x12	15x12	upconv5+pr6+conv5b
pr5	3x3/1	512/1	15x12	15x12	iconv5
upconv4	4x4/2	512/256	15x12	30x24	iconv5
iconv4	3x3/1	769/256	30x24	30x24	upconv4+pr5+conv5b
pr4	3x3/1	256/1	30x24	30x24	iconv4
upconv3	4x4/2	256/128	30x24	60x48	iconv4
iconv3	3x3/1	385/128	60x48	60x48	upconv3+pr4+conv3b
pr3	3x3/1	128/1	60x48	60x48	iconv3
upconv2	4x4/2	128/64	60x48	120x96	iconv3
iconv2	3x3/1	193/64	120x96	120x96	upconv2+pr3+conv2
pr2	3x3x/1	64/1	120x96	120x96	iconv2
upconv1	4x4/2	64/32	120x96	240x192	iconv2
iconv1	3x3/1	97/32	120x96	120x96	upconv1+pr2+conv1
pr1	3x3/1	32/1	120x96	120x96	iconv1

[1] Multiple inputs

[2] Concatenated inputs

5.1.2 Classification Network

The classification network computes a matching score between a pair of video ROIs ($I^{p(t)}$ and $I^{p(t+1)}$) and a high-quality frontal still ROI (I^{S}). As shown in Figure 5.1, video representation $\mathbf{r}_v$ and still representation $\mathbf{r}_s = E_s\left(I^S\right)$ are fed into the fully-connected classification network. Still representation $\mathbf{r}_s$ is obtained by applying a similar encoder to the reconstruction network, where it contains ten strided convolutional layers. The representations $\mathbf{r}_v$ and $\mathbf{r}_s$ are then concatenated and fed into a fully-connected network with two hidden layers of 512 and 32 neurons, and the matching score is calculated.

During operations as shown in Figure 5.3, probe video ROIs from a trajectory are matched against the representations of still ROIs that already computed and preserved in the gallery set.

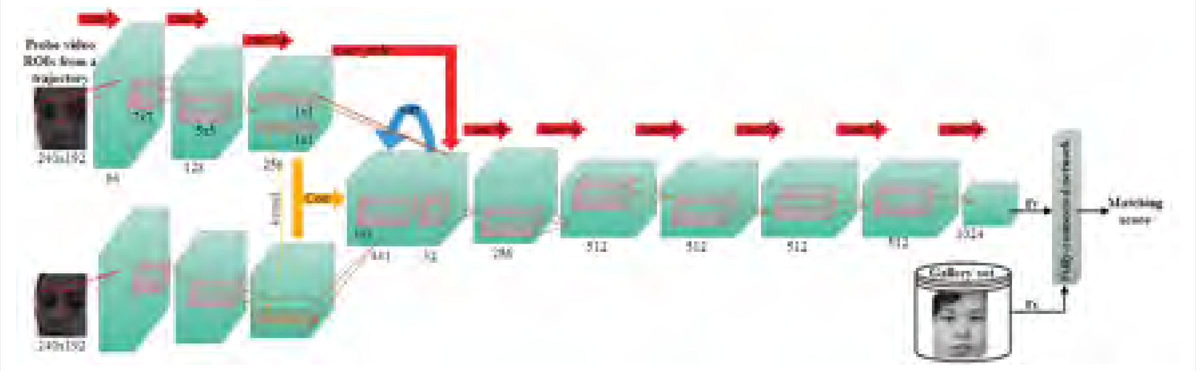


Figure 5.3 The operational phase of the proposed FFA-CNN.

5.1.3 Training FFA-CNN

The goal of training is to design a network that can learn robust representations for still-to-video FR. Given a pair of video ROIs and a single high-quality still ROI, the proposed FFA-CNN is trained through DA for still-to-video FR from a SSPP. To that end, the reconstruction and classification networks are trained through backpropagation using a supervised end-to-end process to achieve identity preserved frontal face reconstruction and accurate still-to-video FR.

During training, parameters of the FFA-CNN are optimized by minimizing a weighted sum of different loss functions including pixel-wise loss L_{pixel}, symmetry loss $L_{symmetry}$, and identity preserving loss $L_{identity}$. For a training set with N batches of $\left\{\left(I^{S}, I^{p(t)}, I^{p(t+1)}\right)\right\}$ and ground-truth target of $\left\{\left(I^{T}, M\right)\right\}$, where I^{T} is the high-quality still corresponding to $I^{p(t)}$ and $I^{p(t+1)}$, and M is the matching labels, respectively, the overall weighted loss function of the FFA-CNN can be formulated as follows:

$$L_{FFA-CNN} = \omega_1 L_{pixel} + \omega_2 L_{symmetry} + \omega_3 L_{identity} \tag{5.1}$$

where ω_i is the weight for the corresponding loss function. The pixel-wise loss L_{pixel} is computed to maintain the consistency of image content at multiple prediction scales (see Figure 5.2).

$$L_{pixel} = \sum_{i=1}^{6} \frac{1}{W_i \times H_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} \left| sub(I^{T})^{i}_{x,y} - (I^{F})^{i}_{x,y} \right| \tag{5.2}$$

where $sub(I^{T})^{i}$ downsamples I^{T} to $(W_i \times H_i)$ using the bilinear subsampling algorithm. Also, i corresponds to prediction layer i in Figure 5.2, where each prediction is generated by applying a convolutional layer to the corresponding upconvolutional output. In addition, symmetry loss $L_{symmetry}$ is considered to sustain the symmetry inherent of faces, where it allows the network to alleviate partial occlusions and large pose variations. It is thereby capable of transferring the appearance of visible part of the face to the missing part.

$$L_{symmetry} = \frac{1}{\frac{W}{2} \times H} \sum_{x=1}^{\frac{W}{2}} \sum_{y=1}^{H} \left| I^{F}_{x,y} - I^{F}_{W-(x-1),y} \right|. \tag{5.3}$$

Finally, identity preserving loss $L_{identity}$ is employed to retain perceptual similarity allowing the capability of accurate FR through an end-to-end learning (Huang *et al.*, 2017). Thus, using the identity preserving loss enforces the network to reconstruct images that can resemble the ground-truth still target (I^T), as well as, to generate discriminant feature representations similar for the same identities using cross-entropy criterion. $L_{identity}$ is exploited in the classification network and defined as follows:

$$L_{identity} = CE(F(\mathbf{r}_s, \mathbf{r}_v), M) \tag{5.4}$$

where CE is the cross-entropy and $(\mathbf{r}_s, \mathbf{r}_v)$ are concatenated to generate a single feature vector. Moreover, $F(\mathbf{r}_s, \mathbf{r}_v)$ is the output of fully-connected classification network that is followed by a softmax activation layer.

Finally, the overall loss function $L_{FFA-CNN}$ is minimized using the stochastic gradient decent with momentum (Adam algorithm). Therefore, the proposed end-to-end multi-task learning framework can simultaneously learn to reconstruct a frontal well-illuminated face, as well as, to compare a pair of video ROIs and the still reference ROI.

5.2 Experimental Results and Discussions

In this section, the proposed FFA-CNN is validated in terms of face reconstruction and feature representations, and compared against the sate-of-the-art CNN-based video-based FR systems.

5.2.1 Experimental Protocol

Validation process of the proposed network is performed considering a random set of 100 unknown subjects including their single stills and some video ROIs for training. Other video ROIs of those subjects from one camera are utilized during evaluation, where their stills are organized as the gallery set and their video ROIs constitute the probe set. In the experiment, the high-quality stills of 100 target individuals (5 different groups of 20 individuals of interest) are randomly selected to participate in the watch-list, while video sequences of the rest along with videos of the target individual are used during operations. In this scenario, one individual at a time is considered as the target individual and appears in the operational scene along with videos of non-target individuals.

For experimental validation, all the still and video ROIs are upscaled to 240x192 pixels to be fed into the proposed framework. The network is implemented using TensorFlow deep learning framework (Allaire *et al.*, 2016) and trained for 60 epochs using Adam algorithm over roughly 90K training samples (45K positives + 45K negatives) by optimizing the overall loss function ($L_{FFA-CNN}$). The positive samples include a pair consecutive video ROIs and a single corresponding still ROI, whereas negative samples include an additional non-corresponding still ROI (I^S). Nevertheless, both stills (I^T) and (I^S) are the same still ROI in positive samples, while the still ROI (I^S) is randomly sampled from other non-target individuals in negative samples. In all of the experiments, the weights of loss function are empirically chosen as $\omega_1 = 0.5$, $\omega_2 = 1$ and $\omega_3 = 1$. In addition, learning rate is set to $1e^{-6}$ and reduced using an exponential method every 1000 iterations with batch size of 10.

The proposed FFA-CNN is compared at transaction-level against ensemble of e-SVMs (Ee-SVMs) (Bashbaghi *et al.*, 2017a), VGG-Face (Parkhi *et al.*, 2015), HaarNet (Parchami *et al.*, 2017a), cross-correlation matching CNN (CCM-CNN) (Parchami *et al.*, 2017b) and CFR-CNN (Parchami *et al.*, 2017c) as the CNN-based state-of-the-art video-based FR methods.

5.2.2 Experimental Results

Figure 5.4 illustrates a random example of ground-truth stills, input probe ROIs and reconstructed face ROIs at different training epochs of the proposed FFA-CNN.

As visualized in Figure 5.4, the network is able to reconstruct still-like frontal faces with neutral expression along different epochs. Accordingly, it can diminish the differences between the target and source domains and generate visually appealing faces. To visually compare the proposed FFA-CNN with the baseline system, a random example of probe ROIs and their corresponding reconstructed ROIs generated by the CFR-CNN (Parchami *et al.*, 2017c) is depicted in Figure 5.5.

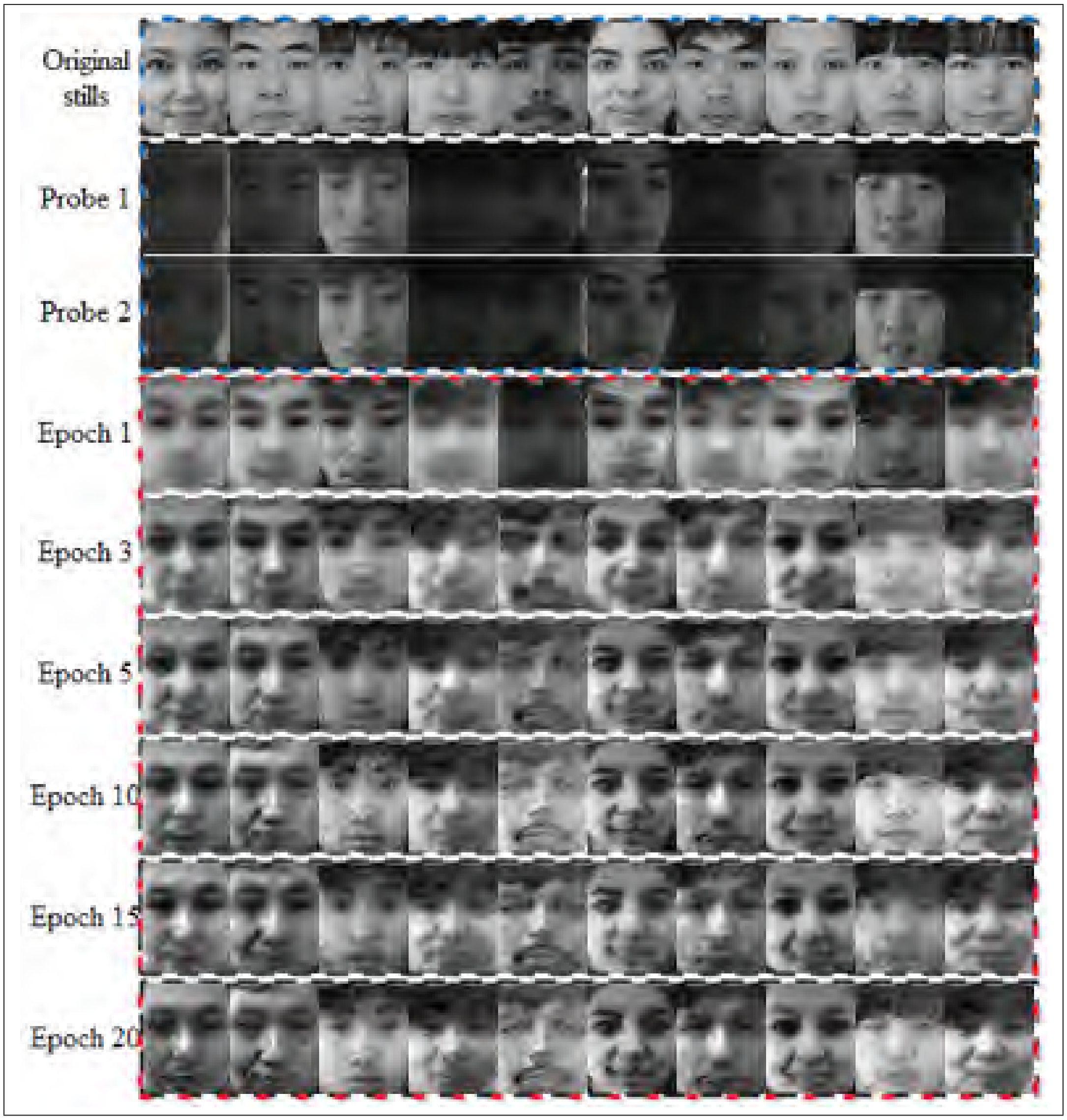


Figure 5.4 A random example of original still and input probe ROIs of random subjects with the corresponding reconstructed faces at different epochs.

As demonstrated in Figures 5.5, ROIs reconstructed by the baseline CFR-CNN (Parchami *et al.*, 2017c) are relatively similar and it fails to reconstruct visually acceptable faces for some probe ROIs. However, the proposed FFA-CNN can reconstruct significantly better faces as shown in Figure 5.4.

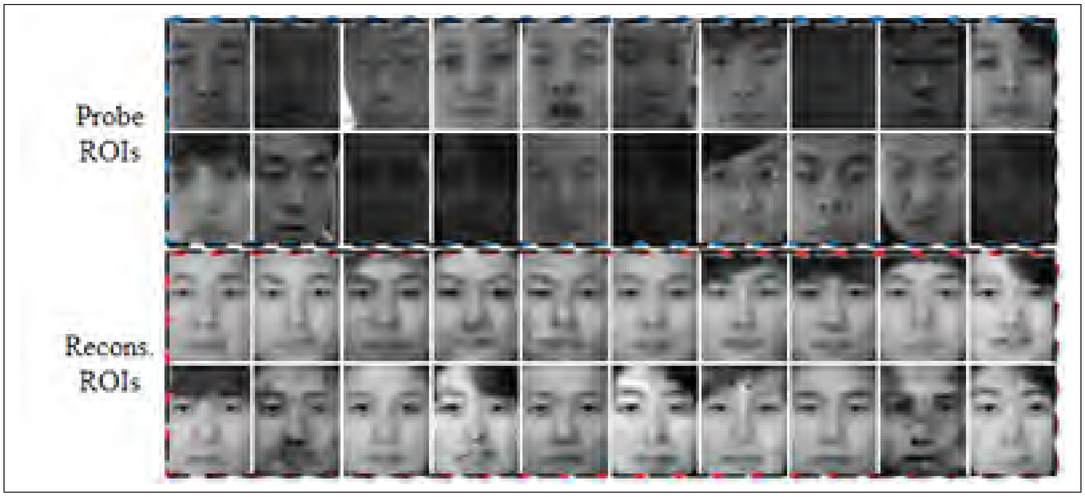


Figure 5.5 A random example of probe ROIs and reconstructed ROIs using the baseline system. The top rows are the probe ROIs and the bottom rows are their corresponding reconstructed face ROIs.

Although the reconstructed faces might not be perceptually separable, but the face embeddings learned by the network through the training process can be utilized for robust video-based FR. In this regard, deep features extracted from the last layer of the network (3x3x1024) for 5 random subjects are mapped on a 2-D space using t-SNE algorithm (Maaten & Hinton, 2008) as shown in Figure 5.6.

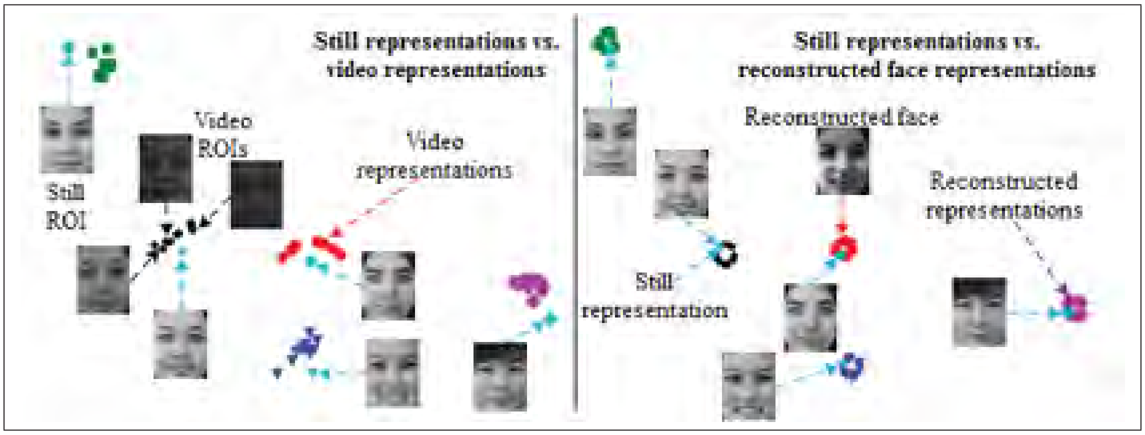


Figure 5.6 Illustration of 2-D feature space of the original stills with input video ROIs (left) and reconstructed face ROIs (right). Subjects are represented by different colors.

As displayed in Figure 5.6 (left), deep feature space of the input video ROIs (video representations) and their corresponding stills are relatively separable, while as illustrated in Figure 5.6 (right) the deep features extracted from the reconstructed face ROIs can be properly classified into different sets w.r.t. their identities. It can be concluded that although the proposed FFA-CNN can generate discriminant representations for the still and video ROIs of the same subjects, but reconstructing frontal faces from input video ROIs can significantly improve the robustness of face representations.

The average AUC and AUPR of the proposed FFA-CNN is compared against the state-of-the-art video-based FR systems over videos of Cox Face DB as shown in Table 5.2.

Table 5.2 Average AUC and AUPR of the proposed network against the CNN-based state-of-the-art systems.

FR systems	**AUC**	**AUPR**
Ee-SVMs Bashbaghi *et al.* (2017a)	98.37±0.65	87.35±1.02
VGG-Face (Parkhi *et al.*, 2015)	64.99±2.49	34.11±2.67
HaarNet (Parchami *et al.*, 2017a)	90.70±2.42	65.64±4.28
CCM-CNN (Parchami *et al.*, 2017b)	89.37±2.12	64.12±3.85
CFR-CNN (Parchami *et al.*, 2017c)	85.55±4.16	53.63±3.57
FFA-CNN (1 rec. branch)	91.59±2.15	58.07±7.62
FFA-CNN (2 rec. branches)	90.45±3.71	57.11±4.40

As presented in Table 5.2, individual-specific Ee-SVMs with HOG face descriptor significantly outperforms the Ee-SVMs using VGG-face deep features. Noted that VGG-Face network receives color images with 224x224x3 pixels as input and returns face descriptors with 4096 features. Thus, COX face images are upscaled and fed into the VGG network, where it generates a feature vector with many zero values (around 400 nonzero values). Since, VGG-Face network was trained on high-quality face images without considering DA constraints, it cannot generate robust features suitable for low-quality video images observed in still-to-video FR. Since faces in COX are grayscale and 48x60 pixels, it is not feasible to fine-tune the VGG network because it requires RGB images with 3 color channels and 224x224 pixels. HaarNet and CCM-CNN provide higher performance, while they require millions of training data and additional fine-tuning process with face synthesizing. They are elegant architectures, where HaarNet is an ensemble of CNNs with a trunk network along with multiple branches and CCM-CNN is a Siamese network that contains 3 identical convolutional branches.

The proposed FFA-CNN is evaluated using both architectures with one and two reconstruction branches as introduced in FlowNetS and FlowNetC (Dosovitskiy *et al.*, 2015), respectively. In FFA-CNN with one reconstruction branch, the reconstruction network is similar to the classification network without correlation and fully-connected layers. However, the proposed FFA-CNN can achieve competitive accuracy with significantly lower number of training data and without any additional fine-tuning. The results also confirms that the proposed FFA-CNN outperforms the baseline autoencoder-based CFR-CNN. The main reason of lower AUPR values in comparing with AUC values is that the operational data is severely imbalanced. Overall, the proposed FFA-CNN can achieve competitive accuracy with significantly lower number of training data and without any additional fine-tuning.

In another experiment, the FFA-CNN is evaluated on the videos of Chokepoint dataset considering 5 individuals of interest. It shows the AUC and AUPR results as 66.58±5.94 and 30.84±5.18, respectively. Thus, it can be concluded that compared to other methods in this chapter, FFA-CNN does not generalized well when tested on different datasets, and suffers from of incompatibilities in scales of images.

As observed in Table 5.2, it can be concluded that the deep learning architectures compared in this chapter cannot perform accurately on highly imbalanced data situation, such as real-world still-to-video FR with SSPP, while individual-specific ensembles can provide a robust solution for such a challenging task. It is worth noting that, CNN-based FR systems can be considered as a more generic solution that can be appropriately exploited for various still-to-still FR tasks, such as face verification.

The complexity of the proposed FFA-CNN is also compared in Table 5.3 against the state-of-the-art video-based FR systems, in terms of number of operations, network parameters, layers and data required for training the network.

Table 5.3 The complexity (number of operations, network parameters, layers and training data) of the state-of-the-art CNN-based FR systems.

FR systems	**Complexity**			
	No. operations	**No. parameters**	**No. layers**	**No. training data**
Ee-SVMs (Bashbaghi *et al.*, 2017a)	2.3M	N_{wl} x 230K	N/A	N_{wl} x 10K
VGG-Face (Parkhi *et al.*, 2015)	31.7B	1.8B	37	2.62M
HaarNet (Parchami *et al.*, 2017a)	3.5B	13.1M	56	1.3M
CCM-CNN (Parchami *et al.*, 2017b)	33.3M	2.4M	30	1.3M
CFR-CNN (Parchami *et al.*, 2017c)	3.7M	1.2M	7	136K
FFA-CNN (1 rec. branch)	3.8B	31.5M	10	90K
FFA-CNN (2 rec. branches)	6.2B	31.5M	10	90K

Collecting adequate training data to train a deep CNN is not feasible in many real-world FR applications, due to the cost and time required to process and label them. It can be observed in Table 5.3 that the proposed FFA-CNN needs only 90 thousands (90K) training data to be properly trained. E-eSVMs was trained on N_{wl}x10K, where N_{wl} is the number of watch-list target individuals. However, VGG-Face, HaarNet and CCM-CNN have been trained on 2.62 million (2.62M), and 1.3 million training samples, respectively. In addition, a complex triplet-based loss function was exploited to train and fine-tune them in order to learn a robust face embedding, where it aims to differentiate between the positive pair of two matching ROI and the negative non-matching ROI.

Overall, the complexity of Ee-SVMs increases linearly w.r.t. the number of watch-list individuals (N_{wl}), and it is not subsequently applicable for a large set of watch-list persons. On the other hands, CNNs are typically complex solutions, while they have a fixed cost regardless of the number of watch-list persons. Although FFA-CNN requires a large number of parameters and operations compared to CCM-CNN and CFR-CNN and it is not specifically helping for the classification/matching, but it can reconstruct visually acceptable faces.

CONCLUSION AND RECOMMENDATIONS

Face recognition has experienced significant advances during the past decades in many applications, such as security and intelligent video surveillance, because it offers remarkable advantages over other biometric modalities (e.g., fingerprint, iris, etc.). In particular, capturing and processing faces in a crowd without any further cooperations of individuals can be a critical issue in different video surveillance scenarios, such as watch-list screening and person re-identification. Nevertheless, designing a face recognition system that can accurately recognize the individuals of interest under unconstrained video surveillance applications is a challenging problem, where faces captured through a network of distributed low-quality video cameras exhibit several variations, such as changes in pose, illumination, facial expression, occlusion and blur. Typically, facial models used for matching against video faces are designed a priori with a limited number of reference face images, while these models are not representative of the operational scene. However, imbalanced data situation encountered in still-to-video face recognition, as well as, differences between the data distributions of still and video domains, magnify the complexity of such real-world applications.

In this thesis, new frameworks were proposed for robust watch-list screening of faces in video surveillance. Different constraints caused by domain adaptation and single sample per person problems were carefully addressed in order to design a reliable still-to-video face recognition system. To that end, several feature extraction techniques, patch-based and random subspace methods were employed to generate multiple face representations due to compensation of limited design samples and providing robustness to intra-class variations. A multi-classifier system was designed specifically for each individual of interest, where specialized 2-class classifier were trained using a single labeled still representation and an abundant of unlabeled video representations. These individual-specific ensemble of classifiers are capable of learning robust facial models during enrollment of a target individual, where they can accurately detect the presence of target individuals during operations under uncontrolled conditions. Dynamic selection and weighting of classifiers were also considered to take the most competent classifiers into account for ensemble fusion. Spatio-temporal recognition can be robustly carried out through aggregating the ensemble scores over trajectories of successive frames. In addition, a deep architecture was proposed through a supervised autoencoder neural network that can be trained end-to-end to reconstruct frontal still-like faces from input video faces using the reconstruction network and to generate robust representations that can be embedded into the fully-connected classification network.

In Chapter 3, several feature extraction techniques and uniform non-overlapping patches were employed to generate a diversified ensemble of SVM classifiers per target individual. Feature extraction techniques were chosen based on their robustness against variety of nuisance factors encountered in video surveillance environments. Since there is only a single reference still per individual of interest during design, videos of unknown non-target individuals in the scene have been used to train the classifiers to exploit the information of operational phase. Hence, videos of background model are more representative of real scene, contrary to other stills in the cohort model. To achieve higher performance, an intuitive dynamic ensemble selection method was proposed to select the most suitable classifiers related to different capture conditions. Finally, accumulating ensemble scores over multiple face captures of corresponding individuals using a high-quality track that were provided by the face tracker significantly improves the overall performance. Extensive experiments with the Chokepoint and COX-S2V video datasets indicated that the integration of patches-based approach and face descriptors into an individual-specific ensemble of e-SVM classifiers provides a significantly higher level of performance than either any single representation, or baseline system containing ensemble of OC-SVMs and template matchers. Results also demonstrated that training a separate classifier for each patch and combining their scores outperforms a single classifier trained using a long feature vector of concatenated patches.

In Chapter 4, individual-specific ensembles of exemplar-SVMs were designed to model a single reference still of target individuals. A novel ensemble-based learning was utilized, where multiple random subspaces were generated for different face descriptors extracted from face patches to effectively provide ensemble diversity and tackle the SSPP constraints. Unlike conventional random subspace methods that completely select the feature subspaces randomly from the entire ROI, semi-random subspaces were exploited to either consider the distribution of face descriptors and to make use of the local spatial relation among each patch. Furthermore, an unsupervised domain adaptation method was used to train the classifiers in the enrollment domain through several training schemes, where video ROIs of non-target individuals were exploited versus a single still ROI to transfer knowledge from the operational domains. Thus, such a system can incorporate knowledge of the operational domains and improve the robustness against several nuisance factors frequently observed in video surveillance operational environments. Two different design scenarios were considered during enrollment to generate a pool of diverse classifiers, where the first scenario exploits additional knowledge acquired from a validation set and a global criterion to select the best subspaces and to construct an efficient system with a compact pool of classifiers. In addition, distance-based dynamic selection and weighting approaches were also proposed to either select or weight the classifiers dynamically during operations. Extensive evidences are provided using the COX-S2V and Chokepoint datasets that the proposed method is effective and comparable against the state-of-the-art methods. Experimental results indicated that integration of the ranked semi-random subspaces into an individual-specific ensembles can yield to a higher level of performance. However, the proposed system with the compact pool of e-SVMs and dynamic weighting can achieve state-of-the-art performance with a significantly lower computational complexity. The results confirmed that all solutions can achieve very high performance at trajectory-level, therefore it is more desirable to focus on low-cost solutions.

Finally, a deep learning-based architecture was proposed in Chapter 5 for robust single sample still-to-video FR using a supervised autoencoder. Still images and videos from the source and target domains were simultaneously considered through a supervised end-to-end training to address the issue of DA. The proposed system contains the reconstruction network that is exploited to create a still-wise frontal face from low-quality and blurry input video probes, as well as, to generate video representations. While, the classification network generates still representations and learns the matching of still and video ROIs. It employed a weighted loss function in the training process that combines pixel-wise loss, symmetry loss and identity preserving loss. It is worth noting that, unsupervised autoencoders cannot perform well, because they are not able to preserve the identity to provide discriminative representations. The experimental results over the challenging COX Face DB demonstrate that the proposed network can effectively apply on video surveillance applications with large variations of face captures. The results also indicate that the proposed network can reconstruct visually convincing faces and generate discriminative representations similar for the same identities. Meanwhile, it can outperform most of the state-of-the-art CNN-based FR systems, while it requires significantly lower number of labeled training data without fine-tuning.

Overall, it can be concluded that conventional methods such as, exploiting multiple handcrafted feature extraction and ensemble-based techniques proposed in Chapter 3 and 4 can appropriately address a specific-purpose solution for still-to-video FR applications, while deep learning methods offer more general-purpose solutions for various FR applications. Currently, deep CNN architectures appear to provide complex solutions that are less suitable for real-time video surveillance applications.

Future Work

Faces captured in video surveillance environments typically suffer from poor quality, due to characteristics of surveillance cameras, as well as, uncontrolled capturing conditions that may lead to variations in ambient lighting, pose, shadowing, and motion blur. Subsequently, the quality of facial captures using surveillance cameras can significantly affect the performance of video-based face recognition systems. Thus, contextual information obtained from faces captured in real-world surveillance environments can be incorporated as a complementary knowledge to either classifier fusion or selection/weighting of classifiers. To that end, different face quality measures could be employed during the operational phase to represent the capturing context. The quality measures (e.g., brightness, contrast, sharpness, illumination, facial pose, etc.), therefore, can be exploited to determine multiple criteria (levels of competence) in order to select or weight the most desirable classifiers dynamically within the pool of classifiers generated during the enrollment phase.

In addition, other biometric modalities or soft biometric traits could be also considered as an additional source of information to build more reliable and accurate face recognition systems suitable for real-world video surveillance applications. To generate diverse face representations, overlapping patches and different resolutions of ROIs can be considered to extract multiple features. However, the focus would be to find a simpler solutions with lower computational complexity that can achieve the sate-of-the-art performance.

Finally, to reconstruct high-quality frontal faces and to jointly learn pose-invariant feature representations through the proposed deep autoencoder network, generative adversarial networks could be also integrated. This type of deep networks can be employed in order to learn generative models via the generator and refinement of reconstruction to achieve discriminative representations using the discriminator. Furthermore, considering the adversarial loss allows low-quality non-frontal video faces to reside in the data distribution of high-quality frontal still faces during the reconstruction process. Meanwhile, recurrent neural networks such as long short-term memory can be exploited because of using their internal memory to process sequences of video frames in order to perform spatio-temporal recognition.

LIST OF PUBLICATIONS

Journal Articles

- Bashbaghi, S., E. Granger, R. Sabourin, and G-A. Bilodeau. Dynamic Ensembles of Exemplar-SVMs for Still-to-Video Face Recognition. *Pattern Recognition*, 69:61-81, 2017.

- Bashbaghi, S., E. Granger, R. Sabourin, and G-A. Bilodeau. Robust Watch-List Screening Using Dynamic Ensembles of SVMs Based on Multiple Face Representations. *Machine Vision and Applications*, 28(1-2):219-241, 2017.

Conference Papers

- Bashbaghi, S., E. Granger, R. Sabourin, and G-A. Bilodeau. Dynamic Selection of Exemplar-SVMs for Watch-List Screening Through Domain Adaptation. *6th International Conference on Pattern Recognition Applications and Methods (ICPRAM'2017)*, Porto, Portugal, February 2017.

- Bashbaghi, S., E. Granger, R. Sabourin, and G-A. Bilodeau. Ensembles of Exemplar-SVMs for Video Face Recognition from a Single Sample Per Person. *12th Advanced Video and Signal Based Surveillance (AVSS'2015)*, Karlsruhe, Germany, August 2015.

- Bashbaghi, S., E. Granger, R. Sabourin, and G-A. Bilodeau. Watch-List Screening Using Ensembles Based on Multiple Face Representations. *22nd International Conference on Pattern Recognition (ICPR'2014)*, Stockholm, Sweden, August 2014.

Book Chapter

- Bashbaghi, S., E. Granger, R. Sabourin, and M. Parchami. Deep Learning Architectures for Face Recognition in Video Surveillance. *Deep learning in Object Detection and Recognition Book*, Springer, 2017.

BIBLIOGRAPHY

Abate, A. F., Nappi, M., Riccio, D. & Sabatino, G. (2007). 2D and 3D face recognition: A survey. *Pattern recognition letters*, 28(14), 1885–1906.

Ahonen, T., Rahtu, E., Ojansivu, V. & Heikkila, J. (2008). Recognition of blurred faces using local phase quantization. *ICPR*, pp. 1-4.

Ahonen, T., Hadid, A. & Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. *Pattern analysis and machine intelligence, IEEE transactions on*, 28(12), 2037–2041.

Allaire, J., Eddelbuettel, D., Golding, N. & Tang, Y. (2016). tensorflow: R Interface to TensorFlow. Consulted at https://github.com/rstudio/tensorflow.

Amira, A. & Farrell, P. (2005). An automatic face recognition system based on wavelet transforms. *Circuits and systems, ISCAS, IEEE international symposium on*, pp. 6252–6255.

Baktashmotlagh, M., Harandi, M., Lovell, B. & Salzmann, M. (2013). Unsupervised domain adaptation by domain invariant projection. *Computer vision (ICCV), IEEE international conference on*, pp. 769-776.

Bansal, A., Castillo, C., Ranjan, R. & Chellappa, R. (2017). The do's and don'ts for cnn-based face verification. *arxiv preprint arxiv:1705.07426*.

Barr, J. R., Bowyer, K. W., Flynn, P. J. & Biswas, S. (2012). Face recognition from video: A review. *International journal of pattern recognition and artificial intelligence*, 26(05).

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*, pp. 1–6.

Bashbaghi, S., Granger, E., Sabourin, R. & Bilodeau, G.-A. (2014). Watch-list screening using ensembles based on multiple face representations. *ICPR*, pp. 4489-4494.

Bashbaghi, S., Granger, E., Sabourin, R. & Bilodeau, G.-A. (2017a). Robust watch-list screening using dynamic ensembles of svms based on multiple face representations. *Machine vision and applications*, 28(1), 219–241.

Bashbaghi, S., Granger, E., Sabourin, R. & Bilodeau, G.-A. (2017b). 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. (2017c). Dynamic selection of exemplar-svms for watch-list screening through domain adaptation. *ICPRAM*.

Batuwita, R. & Palade, V. (2010). Fsvm-cil: fuzzy support vector machines for class imbalance learning. *Fuzzy systems, IEEE transactions on*, 18(3), 558–571.

Bengio, S. & Mariéthoz, J. (2007). Biometric person authentication is a multiple classifier problem. In *Multiple Classifier Systems* (pp. 513–522). Springer.

Bereta, M., Pedrycz, W. & Reformat, M. (2013). Local descriptors and similarity measures for frontal face recognition: A comparative analysis. *Journal of visual communication and image representation*, 24(8), 1213–1231.

Best-Rowden, L., Klare, B., Klontz, J. & Jain, A. K. (2013). Video-to-video face matching: Establishing a baseline for unconstrained face recognition. *Biometrics: Theory, applications and systems (BTAS), IEEE sixth international conference on*, pp. 1–8.

Beveridge, J. R., Phillips, P. J., Bolme, D. S., Draper, B. A., Givens, G. H., Lui, Y. M., Teli, M. N., Zhang, H., Scruggs, W. T., Bowyer, K. W., Flynn, P. J. & Cheng, S. (2013). The challenge of face recognition from digital point-and-shoot cameras. *Biometrics: Theory, applications and systems (BTAS), IEEE sixth international conference on*, pp. 1-8.

Blitzer, J., McDonald, R. & Pereira, F. (2006). Domain adaptation with structural correspondence learning. *Proceedings of the 2006 conference on empirical methods in natural language processing*, (EMNLP '06), 120–128.

Britto, A. S., Sabourin, R. & Oliveira, L. E. (2014). Dynamic selection of classifiers - a comprehensive review. *Pattern recognition*, 47(11), 3665 - 3680.

Canziani, A., Paszke, A. & Culurciello, E. (2016). An analysis of deep neural network models for practical applications. *arxiv preprint arxiv:1605.07678*.

Caruana, R., Munson, A. & Niculescu-Mizil, A. (2006). Getting the most out of ensemble selection. *Icdm*.

Cavalin, P. R., Sabourin, R. & Suen, C. Y. (2012). Logid: An adaptive framework combining local and global incremental learning for dynamic selection of ensembles of hmms. *Pattern recognition*, 45(9), 3544 - 3556.

Cavalin, P., Sabourin, R. & Suen, C. (2013). Dynamic selection approaches for multiple classifier systems. *Neural computing and applications*, 22(3-4), 673-688.

Chang, C.-C. & Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3), 27.

Chawla, N. & Bowyer, K. (2005). Random subspaces and subsampling for 2D face recognition. *CVPR*.

Chellappa, R., Sinha, P. & Phillips, P. J. (2010). Face recognition by computers and humans. *Computer*, 43(2), 46–55.

Chellappa, R., Chen, J., Ranjan, R., Sankaranarayanan, S., Kumar, A., Patel, V. M. & Castillo, C. D. (2016). Towards the design of an end-to-end automated system for image and video-based recognition. *Corr*, abs/1601.07883.

Chen, C., Dantcheva, A. & Ross, A. (2015). An ensemble of patch-based subspaces for makeup-robust face recognition. *Information fusion*, 1-13.

Chen, X., Wang, C., Xiao, B. & Zhang, C. (2014). Still-to-video face recognition via weighted scenario oriented discriminant analysis. *IJCB*.

Cheplygina, V. & Tax, D. (2011). Pruned random subspace method for one-class classifiers. In *Multiple Classifier Systems* (vol. 6713).

Connaughton, R., Bowyer, K. W. & Flynn, P. J. (2013). Fusion of face and iris biometrics. In *Handbook of Iris Recognition* (pp. 219–237). Springer.

Cruz, R. M., Sabourin, R., Cavalcanti, G. D. & Ren, T. I. (2015). Meta-des: A dynamic ensemble selection framework using meta-learning. *Pattern recognition*, 48(5), 1925 - 1935.

Cruz, R. M., Sabourin, R. & Cavalcanti, G. D. (2017). Meta-des.oracle: Meta-learning and feature selection for dynamic ensemble selection. *Information fusion*, 38, 84 - 103.

De la Torre Gomerra, M., Granger, E., Radtke, P. V., Sabourin, R. & Gorodnichy, D. O. (2015). Partially-supervised learning from facial trajectories for face recognition in video surveillance. *Information fusion*, 24(0), 31–53.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. mach. learn. res.*, 7, 1–30.

Deng, W., Hu, J. & Guo, J. (2012). Extended src: Undersampled face recognition via intraclass variant dictionary. *Pami, IEEE trans on*, 34(9), 1864-1870.

Deniz, O., Bueno, G., Salido, J. & la Torre, F. D. (2011). Face recognition using histograms of oriented gradients. *Pattern recognition letters*, 32(12), 1598 - 1603.

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.

Didaci, L., Giacinto, G., Roli, F. & Marcialis, G. L. (2005). A study on the performances of dynamic classifier selection based on local accuracy estimation. *Pattern recognition*, 38(11), 2188 - 2191.

Ding, C. & Tao, D. (2017). Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE trans on PAMI*, PP(99), 1-14. doi: 10.1109/TPAMI.2017.2700390.

Dos Santos, E. M., Sabourin, R. & Maupin, P. (2008). A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. *Pattern recognition*, 41(10), 2993 - 3009.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D. & Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. *ICCV*.

Dreuw, P., Steingrube, P., Hanselmann, H., Ney, H. & Aachen, G. (2009). Surf-face: Face recognition under viewpoint consistency constraints. *BMVC*, pp. 1–11.

Ekenel, H. K., Stallkamp, J. & Stiefelhagen, R. (2010). A video-based door monitoring system using local appearance-based face models. *Computer vision and image understanding*, 114(5), 596–608.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *Systems, man, and cybernetics, part c: Applications and reviews, IEEE trans on*, 42(4), 463-484.

Galar, M., Fernandez, A., Barrenechea, E. & Herrera, F. (2013). Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern recognition*, 46(12), 3460 - 3471.

Galar, M., Fernandez, A., Barrenechea, E. & Herrera, F. (2015). Drcw-ovo: Distance-based relative competence weighting combination for one-vs-one strategy in multi-class problems. *Pattern recognition*, 48(1), 28 - 42.

Gao, S., Zhang, Y., Jia, K., Lu, J. & Zhang, Y. (2015). Single sample face recognition via learning deep supervised autoencoders. *IEEE transactions on information forensics and security*, 10(10), 2108-2118.

Gao, T. & Koller, D. (2011). Active classification based on value of classifier. In *Advances in Neural Information Processing Systems 24* (pp. 1062–1070). Curran Associates, Inc.

Ghodrati, A., Jia, X., Pedersoli, M. & Tuytelaars, T. (2016). Towards automatic image editing: Learning to see another you. In *BMVC*.

Glorot, X., Bordes, A. & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. *Proceedings of the 28th international conference on machine learning (ICML)*, pp. 513–520.

Gong, B., Grauman, K. & Sha, F. (2013). Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. *Proceedings of the 30th international conference on machine learning*, pp. 222–230.

Gopalan, R., Li, R. & Chellappa, R. (2011). Domain adaptation for object recognition: An unsupervised approach. *Computer vision (ICCV), IEEE international conference on*, pp. 999-1006.

Granger, E., Khreich, W., Sabourin, R. & Gorodnichy, D. O. (2012). Fusion of biometric systems using boolean combination: an application to iris–based authentication. *International journal of biometrics*, 4(3), 291–315.

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

He, X., Yan, S., Hu, Y. & Zhang, H.-J. (2003). Learning a locality preserving subspace for visual recognition. *Computer vision, ninth IEEE international conference on*, pp. 385–392.

Hong, S., Im, W., Ryu, J. & Yang, H. S. (2017). Sspp-dan: Deep domain adaptation network for face recognition with single sample per person. *arxiv preprint arxiv:1702.04069*.

Hu, J. (2016). Discriminative transfer learning with sparsity regularization for single-sample face recognition. *Image and vision computing*.

Huang, G. B., Lee, H. & Learned-Miller, E. (2012). Learning hierarchical representations for face verification with convolutional deep belief networks. *CVPR*.

Huang, R., Zhang, S., Li, T. & He, R. (2017). Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arxiv preprint arxiv:1704.04086*.

Huang, Z., Shan, S., Zhang, H., Lao, S., Kuerban, A. & Chen, X. (2013a). Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on cox-s2v dataset. In *Computer Vision–ACCV* (pp. 589–600). Springer.

Huang, Z., Zhao, X., Shan, S., Wang, R. & Chen, X. (2013b). Coupling alignments with recognition for still-to-video face recognition. *Computer vision (ICCV), IEEE international conference on*, pp. 3296–3303.

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. *IP, IEEE trans on*, 24(12), 5967-5981.

Imam, T., Ting, K. M. & Kamruzzaman, J. (2006). z-svm: an svm for improved classification of imbalanced data. In *AI: Advances in Artificial Intelligence* (pp. 264–273). Springer.

Jain, A. K. & Ross, A. (2002). Learning user-specific parameters in a multibiometric system. *Image processing, proceedings, international conference on*, 1, 1–57.

Jain, A. K., Ross, A. & Prabhakar, S. (2004). An introduction to biometric recognition. *Circuits and systems for video technology, IEEE transactions on*, 14(1), 4–20.

Juneja, M., Vedaldi, A., Jawahar, C. & Zisserman, A. (2013). Blocks that shout: Distinctive parts for scene classification. *Computer vision and pattern recognition (CVPR), IEEE conference on*, pp. 923-930.

Kamgar-Parsi, B., Lawson, W. & Kamgar-Parsi, B. (2011). Toward development of a face recognition system for watchlist surveillance. *PAMI, IEEE trans on*, 33(10), 1925-1937.

Kan, M., Shan, S., Su, Y., Xu, D. & Chen, X. (2013). Adaptive discriminant learning for face recognition. *Pattern recognition*, 46(9), 2497–2509.

Kan, M., Shan, S., Chang, H. & Chen, X. (2014). Stacked progressive auto-encoders (spae) for face recognition across poses. *CVPR*.

Kemmler, M., Rodner, E., Wacker, E.-S. & Denzler, J. (2013). One-class classification with gaussian processes. *Pattern recognition*, 46(12), 3507 - 3518.

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M. & Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark a. *IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1931-1939.

Ko, A. H., Sabourin, R., Britto, A. S. & Jr. (2008). From dynamic classifier selection to dynamic ensemble selection. *Pattern recognition*, 41(5), 1718 - 1731.

Kotsiantis, S. (2011). Combining bagging, boosting, rotation forest and random subspace methods. *Artificial intelligence review*, 35(3), 223-240.

Krawczyk, B. & Cyganek, B. (2015). Selecting locally specialised classifiers for one-class classification ensembles. *Pattern analysis and applications*, 1-13.

Krawczyk, B. & Wozniak, M. (2014). Diversity measures for one-class classifier ensembles. *Neurocomputing*, 126(0), 36 - 44.

Krawczyk, B., Wozniak, M. & Cyganek, B. (2014). Clustering-based ensembles for one-class classification. *Information sciences*, 264(0), 182 - 195.

Kuncheva, L. & Whitaker, C. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2), 181-207.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. *ICASSP*.

Li, F. & Wechsler, H. (2005). Open set face recognition using transduction. *Pattern analysis and machine intelligence, IEEE transactions on*, 27(11), 1686–1697.

Li, Q., Yang, B., Li, Y., Deng, N. & Jing, L. (2013a). Constructing support vector machine ensemble with segmentation for imbalanced datasets. *Neural computing and applications*, 22(1), 249–256.

Li, W., Duan, L., Xu, D. & Tsang, I. (2014). Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *PAMI, IEEE trans on*, 36(6), 1134-1148.

Li, Y., Shen, W., Shi, X. & Zhang, Z. (2013b). Ensemble of randomized linear discriminant analysis for face recognition with single sample per person. *Automatic face and gesture recognition (FG), 10th IEEE international conference and workshops on*, pp. 1–8.

Liao, S., Jain, A. K. & Li, S. Z. (2013). Partial face recognition: Alignment-free approach. *Pattern analysis and machine intelligence, IEEE transactions on*, 35(5), 1193–1205.

Liu, C. & Wechsler, H. (2002). Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *Image processing, IEEE transactions on*, 11(4), 467–476.

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

Ma, A., Li, J., Yuen, P. & Li, P. (2015). Cross-domain person reidentification using domain adaptation ranking svms. *IP, IEEE trans on*, 24(5), 1599-1613.

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

Malisiewicz, T., Gupta, A. & Efros, A. (2011a). Ensemble of exemplar-svms for object detection and beyond. *ICCV*.

Malisiewicz, T., Gupta, A. & Efros, A. A. (2011b). Ensemble of exemplar-svms for object detection and beyond. *Computer vision (ICCV), IEEE international conference on*, pp. 89–96.

Matikainen, P., Sukthankar, R. & Hebert, M. (2012). Classifier ensemble recommendation. In *ECCV, Workshops and Demonstrations*. Springer Berlin Heidelberg.

Matta, F. (2008). Video person recognition strategies using head motion and facial appearance. *University of nice sophia-antipolis*.

Matta, F. & Dugelay, J.-L. (2009). Person recognition using facial video information: A state of the art. *Journal of visual languages and computing*, 20(3), 180 - 187.

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A. & Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *CVPR*.

Misra, I., Shrivastava, A. & Hebert, M. (2014). Data-driven exemplar model selection. *Applications of computer vision (WACV), IEEE winter conference on*, pp. 339-346.

Mokhayeri, F., Granger, E. & Bilodeau, G. A. (2015). Synthetic face generation under various operational conditions in video surveillance. *IEEE international conference on image processing (ICIP)*, pp. 4052-4056.

Nourbakhsh, F., Granger, E. & Fumera, G. (2016). An extended sparse classification framework for domain adaptation in video surveillance. *ACCV, workshop on human identification for surveillance*, pp. 360–376.

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

Pagano, C., Granger, E., Sabourin, R. & Gorodnichy, D. O. (2012). Detector ensembles for face recognition in video surveillance. *Neural networks (IJCNN), the international joint conference on*, pp. 1–8.

Pan, S. J. & Yang, Q. (2010). A survey on transfer learning. *Kde, IEEE trans on*, 22(10), 1345-1359.

Pan, S. J., Kwok, J. T. & Yang, Q. (2008). Transfer learning via dimensionality reduction. *Proceedings of the 23rd national conference on artificial intelligence - volume 2*, (AAAI'08), 677–682.

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

Parchami, M., Bashbaghi, S. & Granger, E. (2017b). 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. & 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*.

Patel, V., Gopalan, R., Li, R. & Chellappa, R. (2015). Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3), 53-69.

Qiu, Q., Ni, J. & Chellappa, R. (2014). Dictionary-based domain adaptation for the re-identification of faces. In *Person Re-Identification, Advances in Computer Vision and Pattern Recognition* (pp. 269-285).

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press.

Roweis, S. T. & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2323–2326.

Scholkopf, B. & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

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

Shaokang, C., Sandra, M., Mehrtash T, H., Conrad, S., Abbas, B., Brian C, L. et al. (2011). Face recognition from still images to video sequences: A local-feature-based framework. *EURASIP journal on image and video processing*, 2011.

Shekhar, S., Patel, V., Nguyen, H. & Chellappa, R. (2013). Generalized domain-adaptive dictionaries. *CVPR*.

Skurichina, M. & Duin, R. P. W. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern analysis and applications*, 5(2), 121-135.

Sun, Y., Wang, X. & Tang, X. (2013). Hybrid deep learning for face verification. *ICCV*.

Sun, Y., Chen, Y., Wang, X. & Tang, X. (2014a). Deep learning face representation by joint identification-verification. In *NIPS*.

Sun, Y., Wang, X. & Tang, X. (2014b). Deep learning face representation from predicting 10,000 classes. *CVPR*.

Sun, Y., Wang, X. & Tang, X. (2015). Deeply learned face representations are sparse, selective, and robust. *CVPR*.

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

Tan, X., Chen, S., Zhou, Z.-H. & Zhang, F. (2006). Face recognition from a single image per person: A survey. *Pattern recognition*, 39(9), 1725–1745.

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

Veropoulos, K., Campbell, C., Cristianini, N. et al. (1999). Controlling the sensitivity of support vector machines. *Proceedings of the international joint conference on artificial intelligence*, 1999, 55–60.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 11, 3371–3408.

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

Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H. & Ma, Y. (2012). Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *Pami, IEEE trans on*, 34(2), 372-386.

Wang, C. & Mahadevan, S. (2009). Manifold alignment without correspondence. *Proceedings of the 21st international jont conference on artifical intelligence*, (IJCAI'09), 1273–1278.

Wang, H., Liu, C. & Ding, X. (2015). Still-to-video face recognition in unconstrained environments. *Proc. SPIE, image processing: Machine vision applications*.

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. *Computer vision and pattern recognition workshops (CVPRW), IEEE computer society conference on*, pp. 74–81.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S. & Ma, Y. (2009). Robust face recognition via sparse representation. *Pattern analysis and machine intelligence, IEEE transactions on*, 31(2), 210–227.

Xie, C., Kumar, B. V., Palanivel, S. & Yegnanarayana, B. (2004). A still-to-video face verification system using advanced correlation filters. In *Biometric Authentication* (pp. 102–108). Springer.

Yang, J., Yan, R. & Hauptmann, A. G. (2007). Cross-domain video concept detection using adaptive svms. *Proceedings of the 15th international conference on multimedia*, (MULTIMEDIA '07), 188–197.

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.

Yim, J., Jung, H., Yoo, B., Choi, C., Park, D. & Kim, J. (2015). Rotating your face using multi-task deep neural network. *CVPR*.

Zeng, Z.-Q. & Gao, J. (2009). Improving svm classification with imbalance data set. *Neural information processing*, pp. 389–398.

Zhang, J., Yan, Y. & Lades, M. (1997). Face recognition: eigenface, elastic matching, and neural nets. *Proceedings of the IEEE*, 85(9), 1423–1435.

Zhang, Y. & Wang, D. (2013). A cost-sensitive ensemble method for class-imbalanced datasets. *Abstract and applied analysis*, 2013.

Zhang, Y. & Martínez, A. M. (2004). From stills to video: Face recognition using a probabilistic approach. *Computer vision and pattern recognition workshop, CVPRW, conference on*, pp. 78–78.

Zhao, W., Chellappa, R., Phillips, P. J. & Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4), 399–458.

Zhou, S., Krueger, V. & Chellappa, R. (2003). Probabilistic recognition of human faces from video. *Computer vision and image understanding*, 91(1), 214–245.

Zhu, Y., Li, Y., Mu, G., Shan, S. & Guo, G. (2016). Still-to-video face matching using multiple geodesic flows. *Information forensics and security, IEEE trans on*, 11(12), 2866-2875.

Zhu, Y., Liu, J. & Chen, S. (2009). Semi-random subspace method for face recognition. *Image and vision computing*, 27(9), 1358 - 1370.

Zhu, Z., Luo, P., Wang, X. & Tang, X. (2014a). Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*.

Zhu, Z., Luo, P., Wang, X. & Tang, X. (2014b). Recover canonical-view faces in the wild with deep neural networks. *arxiv preprint arxiv:1404.3543*.

Zong, W. & Huang, G.-B. (2011). Face recognition based on extreme learning machine. *Neurocomputing*, 74(16), 2541 - 2551.

Zou, J., Ji, Q. & Nagy, G. (2007). A comparative study of local matching approach for face recognition. *Image processing, IEEE transactions on*, 16(10), 2617–2628.