Recognizing People in Images and Videos

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

Doctor of Philosophy
in
Computer Science and Engineering

by

Vijay Kumar
201299710
vijaykumar.r@research.iiit.ac.in

Center for Visual Informational Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
July 2019

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis titled "Recognizing People in Images and Videos" by Vijay Kumar has been carried out under our supervision and is not submitted elsewhere for a degree.

Date	Advisor: Prof. Anoop Namboodiri
Date	Advisor: Prof. C. V. Jawahar

To my parents

Acknowledgments

The successful completion of this thesis has been possible through the assistance of many people. I am forever indebted to all the people who have helped me.

First and foremost, I would like to express my sincere gratitude to my PhD advisors Prof. Anoop Namboodiri and Prof. C. V. Jawahar for their constant support and motivation during my stay at IIITH. They helped in formulating my research statement, provided technical and financial assistance throughout my PhD period. Their constant push for hard work and excellence ensured that I focus on important research problems that are more impactful.

I would like to extend my sincere thanks to my PhD committee members Prof. Subhasis Chaudhuri, Prof. Vineet Balasubramanian, and Dr. Ramanathan Subramanian for their valuable comments and suggestions on this thesis draft. Many thanks to Dr. Sumohana Channappayya for agreeing to be part of my thesis defense committee.

I thank the CVIT staff - Siva, Satya, Rajan and Nandini for helping me on travel grants, dataset annotations and various administrative activities. The generous PhD fellowship from TCS provided me financial stability for which I am forever grateful.

I am grateful to dearest friends and colleagues in the CVIT - Pritish, Aniket, Arunava, Saurabh, Praveen, Suriya, Aditya, Avijit, Rajvi, Minesh, Jobin, Anand, Nataraj, Yashaswi and many others for their friendship over these years. Our regular fun conversations, activities and dinner outings ensured that I remain sane during the stressful times and deadlines. The informal discussions and study group sessions we had improved my technical capabilities immensely. I am also thankful to my other IIIT friends Falak, Harit, Lipika, Ajinkya, Neeraj for their friendship over the years.

I am thankful to undergraduate friends - Shishir, Chetan, Gurudutt, Rakesh, Rohan, Pruthvi, Palaksh and Vikrant for their constant support and motivation. I am thankful to Madhu for being there for me. I have no words to express my gratitude to my parents for their unconditional love, support and sacrifice. I am thankful to my sisters (Anitha and Sangeetha), brother-in-laws and uncle for being there for me.

Abstract

Cameras and mobile phones have become integral part of our everyday lives as they become portable, powerful and cheaper. We capture and share hundreds of pictures and videos with our friends, family and social connections. Similarly, large volume of such visual content is generated in surveillance, entertainment, and biometrics applications. Without any doubt, people are the most important objects that dominate in these visual content. For instance, photos taken in a family event or movie videos focus around humans. It is utmost important to automatically detect, identify and analyze people appearing in images to obtain a better understanding of these content and make decisions around them.

In this thesis, we consider the problem of person detection and recognition in images. This is a well explored topic in vision community with a vast literature focused on these problems. The current state-of-the-art recognition systems are able to identify people with high degree of accuracy in scenarios where images have high resolution, contain visible and near frontal faces, and recognition systems have access to sufficiently large training gallery. However, these systems need significant improvement in challenging real-world applications such as surveillance or entertainment videos where one needs to handle several practical issues such as non-visibility of faces, limitation of training samples, domain mismatch, etc in addition to other instance variations such as pose, illumination, and resolution. While there are plenty of challenges pertaining to person recognition, we are interested in some of the open challenges that are relevant from the deployment perspective in diverse recognition scenarios.

We first consider people detection that is a pre-requisite for a recognition system. We detect people in images by detecting their faces through an exemplar based detector. Exemplar approach detects faces through hough voting using an exemplar training database indexed with bag-of-words method. We propose two key ideas referred as "Visual phrases" and "Contextual weighting" into the exemplar approach that improves its performance significantly. We show that visual phrases which encode dependencies between visual features are discriminative and propose a strategy to incorporate them into exemplar voting. We also introduce the notion of spatial consistency for a visual feature which weights each oc-

currence of a feature based on its global context. Our evaluation of popular in-the-wild face detection benchmarks demonstrate significant improvement obtained using these proposed ideas.

We then focus on person recognition and consider several issues encountered in practical recognition systems. We initially address the common and important issue of "unavailability of sufficient training samples" during recognition. We propose a solution based on semi-supervised learning that can efficiently learn from a small amount of labeled data and a large amount of unlabeled data. We demonstrate how the similarities between labeled and unlabeled samples can be effectively exploited to improve the performance. We then consider the problem of "domain mismatch" between training gallery (source) and probe instances (target). We consider a recognition setup in which the objective is to identify people in a collection of probe images using a training gallery collected from different domain. We propose a novel two-stage solution that generates the labels of a few confident seed images from the target domain and propagate their labels to remaining images using a graph based framework. We evaluate our approach in several practical recognition scenarios such as movie videos and photo-albums.

We then consider a different recognition scenario in which faces are not completely visible due to occlusion or people facing away from the camera. To deal with such occluded and partially or completely non-visible faces, we exploit information from other body regions such as head, upper body and body to improve the recognition. When considering different body regions, pose of different body regions pose a serious challenge. To handle the issues of unreliable facial region and pose variation, we propose a technique that learns multiple pose-specific representations from different body regions. Our approach involves training a separate deep convolutional network for each pose and then combining their predictions using adaptive weights determined by the pose of the person.

Person recognition approaches based on multiple body regions however require training multiple deep convolution networks for different body regions resulting in large number of parameters with slower training and testing procedures. To overcome these, we develop an end-to-end person recognition approach based on pooling and aggregation of discriminative features from multiple body regions. Our end-to-end convolutional network pools features from several pre-determined region of interests and adaptively aggregates them using an attention mechanism to produce a compact representation. We evaluate our single end-to-end trained model on multiple person recognition benchmarks and show its effectiveness over multiple models trained on different body regions.

We finally note that all of our work is developed with a keen focus on their applicability in real world applications. We have created and publicly released datasets and source code during the process.

Contents

List of Figures

List of Tables

Publications

Publications related to the thesis:

Manuscript under review:

1. *Vijay Kumar*, Anoop Namboodiri, C. V. Jawahar, Region Pooling and Adaptive Feature Aggregation for End-to-End Person Recognition

Journal:

2. *Vijay Kumar*, Anoop Namboodiri, C. V. Jawahar, Semi-supervised Annotation of Faces in Image Collection, Signal, Image and Video Processing, Springer, 2017.

Conference:

3. *Vijay Kumar*, Anoop Namboodiri, Manohar Paluri, C. V. Jawahar, Pose-Aware Person Recognition, International Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

4. *Vijay Kumar*, Anoop Namboodiri, C. V. Jawahar, Visual Phrases for Exemplar Face Detection, International Conference on Computer Vision (ICCV) Santiago, Chile, 2015.

5. *Vijay Kumar*, Anoop Namboodiri, C. V. Jawahar, Face Recognition in Videos by Label Propagation, International Conference on Pattern Recognition (ICPR) Stockholm, Sweden, 2014.

6. Shankar Setty, Moula Husain, Parisa Beham, Jyothi Gudavalli, Menaka Kandasamy, Radhesyam Vaddi, Vidyagouri Hemadri, J C Karure, Raja Raju, *Vijay Kumar*, C. V. Jawahar, Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations, NCVPRIPG, Jodhpur, India, 2013.

7. *Vijay Kumar*, Anoop Namboodiri, C. V. Jawahar, Sparse Representation based Face Recognition with Limited Labeled Samples, Second Asian Conference on Pattern Recognition (ACPR), Okinawa, Japan, 2013.

Other publications:

1. Himangi Saraogi, Rahul Anand Sharma, *Vijay Kumar*, Event Recognition in Broadcast Soccer Videos, Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP) Guwahati, India, 2016.

2. *Vijay Kumar*, Raghavendra R, Anoop Namboodiri, Christoph Busch, Robust Transgender Face Recognition: Approach based on Appearance and Therapy Factors, International Conference on Identity, Security and Behavior Analysis (ISBA) Sendai, Japan, 2016.

3. Hiba Ashan, *Vijay Kumar*, C. V. Jawahar, Multi-label Annotation of Music, International Conference on Advances in Pattern Recognition (ICAPR) Kolkata, India, 2015.

4. *Vijay Kumar*, Harit Pandya, C. V. Jawahar, Identifying Ragas in Indian Music, International Conference on Pattern Recognition (ICPR), Stockholm, Sweden, 2014.

5. Vijayendra G, *Vijay Kumar*, Sanjay Rawat, Category based Malware Detection in Android, International Symposium on Security in Computing and Communications (SSCC), New Delhi, India, 2014.

6. *Vijay Kumar*, Amit Bansal, Gautam Tulsiyan, Anand Mishra, Anoop Namboodiri, C. V. Jawahar, Sparse Document Coding for Image Restoration, International Conference on Document Analysis and Recognition (ICDAR), Washington D.C, USA 2013.

Chapter 1

Introduction

We live in an era of data overload. There has been an explosion in the amount of visual content such as images, videos, social media content and blogs generated. For instance, it is estimated that more than 100 hours of video is being uploaded in the YouTube every minute, more than 300 million photos are uploaded in Facebook everyday, and more than 2 million blogs are written worldwide everyday. Apart from social medium, large volumes of data are collected in surveillance and authentication systems, autonomous driving, wearable computers, *etc*. This sudden data outburst can be attributed to technological advances, that have enabled us to acquire data at low cost using cameras and smart phones, and process such humungous data with powerful hardware. In order to make sense of such large volume of redundant data, it is necessary to develop intelligent algorithms that extract meaningful information and infer higher level decisions, the way humans do.

Humans are ubiquitous in images and videos, and are the most important components for a machine's world. As shown in Figure 1.1, humans are the central objects in many systems such as photo albums, entertainment videos and games, movies, sports broadcast videos, surveillance and authentication systems. Thus, one of the major goals of artificial intelligence since its early days is to provide machines the ability to understand and make decisions around humans appearing in the images. The problems associated with humans vary widely from their detection and identification in images to recognizing their expressions or pose. While we can perform all these tasks with ease, it is still very challenging for machines due to wide variations depicted by humans in images. Some of the major variations seen in real-world images are the pose, illumination, occlusion, resolution, camera factors, *etc*. While current day algorithms achieve superior results on restricted settings (people looking towards camera, high resolution images, *etc*.,) they still underperform in a completely unrestricted real-world setting (surveillance and entertainment videos, *etc*.,).

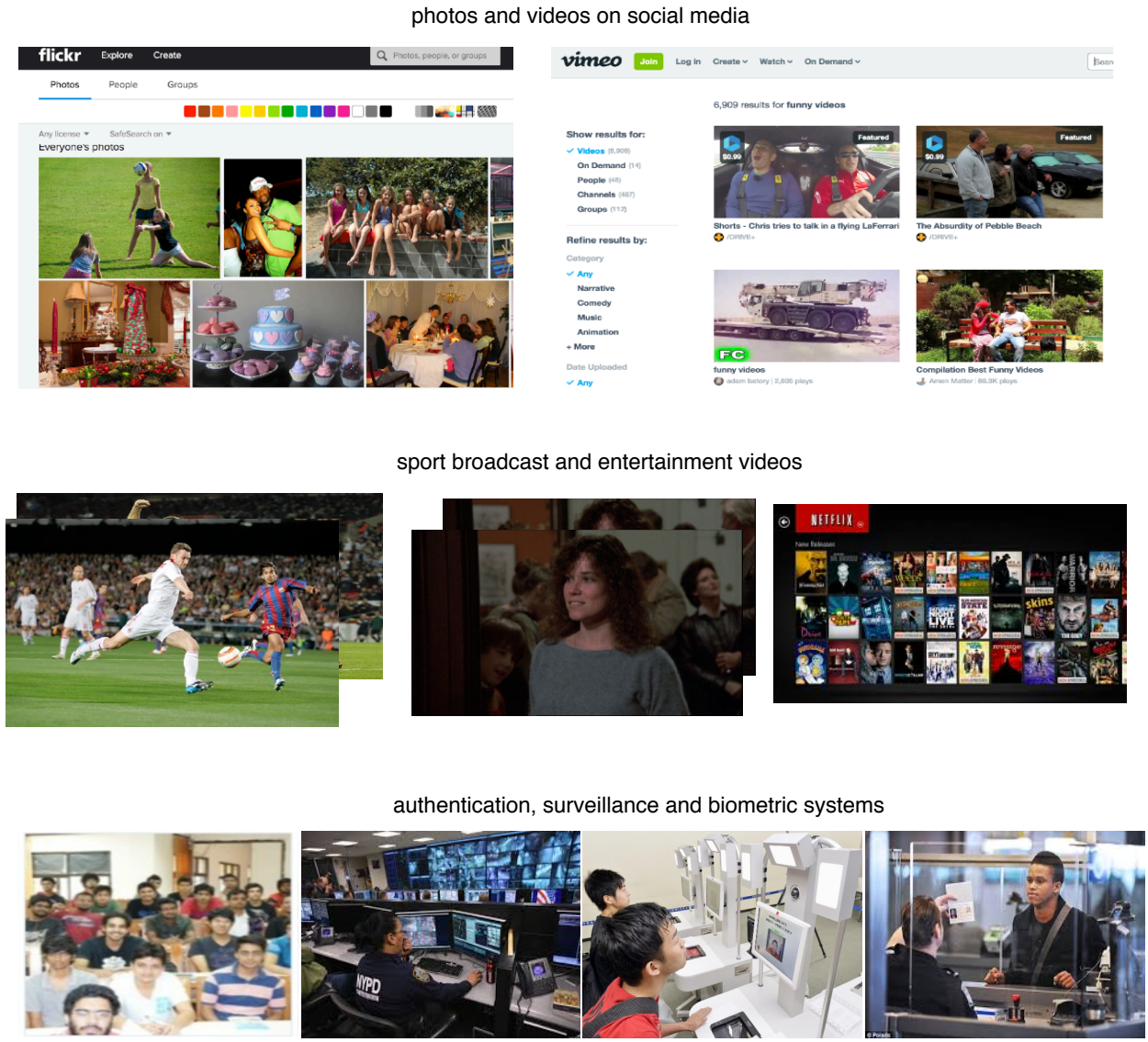


Figure 1.1 Few example systems where humans are the central objects. Identification of people is important to organize albums and videos (top), index and gather higher level insights (middle) and authenticate users (bottom). While the end goal of these applications differ, person recognition is vital in all these applications.

1.1 Problem, Challenges and Motivation

The thesis is particularly focused towards *detection and identification* of people in images and videos. Given an image or a video, our goal is to develop algorithms that can locate people appearing in these images and correctly predict their identities. There are many challenges associated with the problem of person detection and recognition as listed below.

- Appearance variations: Human faces are expressive and body is highly deformable. In addition, the changes in body and head pose, age variations, resolution of the images, occlusions, lighting conditions can cause lot of confusion to the detection and recognition systems. The detection and recognition models should be robust and invariant to these variations in order to achieve high performance.

- Non-visibility of face: Recognizing people from their faces is perhaps the widely used recognition strategy due to its strong discriminative ability. However, in many practical scenarios faces of the person are not completely visible. This may occur due to severe occlusion or subject is facing away from the camera or camera angle. In order to recognize people in such scenarios it may be necessary to look for cues from other body regions.

- Availability of data: Recognition systems predict the identity of subjects by training models on a pre-defined gallery of instances of the subjects that needs to be identified. The performance of recognition systems depend on the amount of available labeled training data. While larger the training data better is the performance, it is not always feasible in many practical applications to possess large gallery of the subjects. The data availability is another factor that should be considered when designing "practical" recognition systems such as biometric or authentication systems, which usually contain one or a few training instances.

- Domain mismatch: In many recognition scenarios it may happen that the training images collected from a particular source have different distribution compared to images observed during testing. An example is automatic identification of actors in movie frames using still-images collected from the Google search. The training and testing instances in such scenarios will have completely different camera, imaging and background conditions, and such differences in the data distribution between source and target instances may degrade the performance if not handled properly.

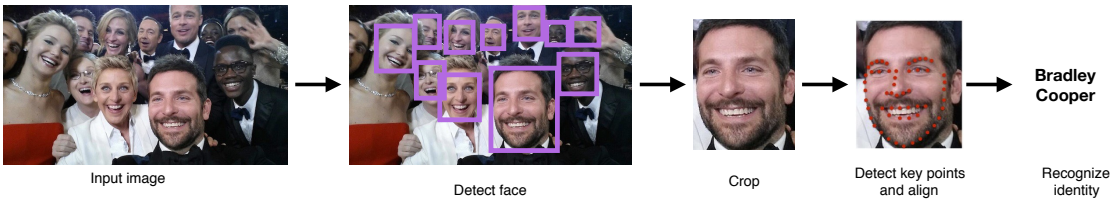


Figure 1.2 A typical recognition pipeline shown with faces. Faces are detected, cropped and aligned before passing through a recognition algorithm that predicts the identities of the subjects appearing in the given image.

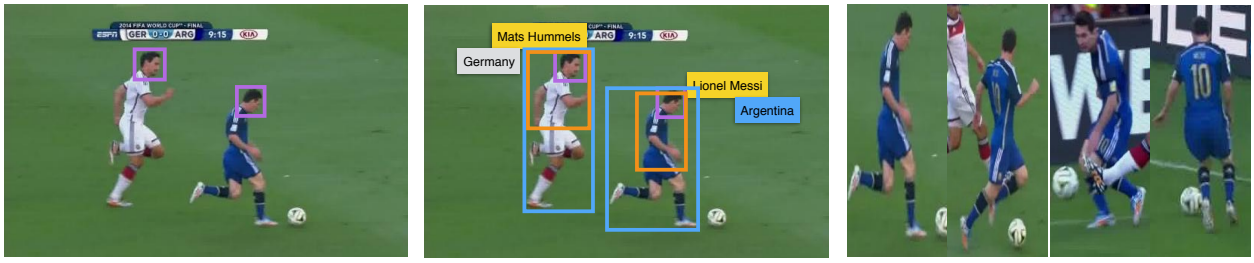
- Complexity: The objective of recognition systems is to build models that work in multiple and diverse conditions. However, the inherent complexity of the domain plays a crucial role in deciding the performance of the systems. For instance, a recognition system that performs well on photo-albums may perform badly when used to identify soccer players in videos or actor identification in movies. It is thus necessary to evaluate algorithms in diverse settings to develop a robust systems that are not biased to particular working conditions.

1.2 Recognition Pipeline

A typical identity recognition pipeline consists of two or more steps as shown in Figure 1.2. Given an image, body regions are first detected. While face based recognition systems are the most popular, one can recognize identity based on multiple body regions such as head, upper body and body in challenging scenarios. In Figure 1.2, we demonstrate the recognition pipeline for faces that is also applicable for other body regions. A face detection algorithm is applied on a given image to obtain the locations of all the face regions appearing in the image. The detection algorithm may make mistakes and produce false positives in this step and practical recognition systems should have the capability to reject such false positives. The cropped faces are then passed through an optional alignment procedure where facial key points are first detected, and then used to align the face with respect to a template. The processed image is then passed on to a recognition algorithm that predicts the identity of the subject.

1.3 Thesis Contributions

With the aforementioned motivations in mind, we focus on several key challenges associated with people recognition in practical scenarios as illustrated in Figure 1.3. We list below the problems at-



(a) Face detection (b) Recognition using facial and multi body cues (c) Recognition using domain cues

Figure 1.3 We are interested in the problem of person recognition. We focus on their (a) detection, and recognition using (b) face and multi-body cues and (c) domain information about the presence of multiple instances of each subject.

tempted in our thesis. All these work fall under the "umbrella" of person recognition. We consider each of these sub-problems independently, and evaluate our methods on different benchmark datasets to allow direct comparison with state-of-the-art algorithms for each problem.

1. **People detection:** This is the first step in recognition pipeline that aims at predicting the location of people in images whose identities have to be recognized. One could either develop algorithms to locate complete body region or just face to infer other body regions later. We consider face detection in our work due to its popularity, wide variety of use-cases and active participation of the community to face detection open challenges. Apart from being a critical precursor to recognition tasks, detection directly benefits several applications such as autofocus in cameras, crowd monitoring, pay-per-face marketing campaigns.

2. **Recognition with limited labeled samples:** We consider people recognition in the presence of limited number of training samples. We show how recognition systems that have access to unlabeled data can leverage it to improve the recognition performance. We focus on semi-supervised learning technique for person recognition to deal with lack of data availability and show their superiority over supervised schemes that only use labeled examples.

3. **Improving recognition with domain information:** We consider a specific problem of people recognition in image collections such as photo-albums and videos, where it is required to identify a set of correlated images simultaneously. In such scenario, we demonstrate how it is beneficial to exploit the domain information and presence of multiple target images to obtain superior performance.

4. **Recognition using multiple body cues:** We focus on recognition algorithms that predict the identity of people using several body regions such as head, upper body, body *etc*, along with face. We propose a state-of-the-art deep convolutional neural network (CNN) architecture to learn pose-aware person representations from multiple body regions. To overcome the redundancy of training multiple CNN models for different body regions, we explore an end-to-end deep CNN architecture wherein a single model extracts and aggregates the features from different body regions in an adaptive fashion.

1.4 Scope of the Thesis

This thesis focuses on person recognition, and few of its issues in practical settings. This is in contrary to vast literature on human identity recognition that focuses on pushing the state-of-the-art on popular verification and identification benchmarks such as LFW [55], Mega Face [98] or YouTube Faces [148]. These works focus primarily on improving the algorithm, and have requirements in terms of availability of large training data, visibility of faces and so on. Our thesis takes a different view on person recognition and looks into some of the issues encountered in practical and deployment scenarios, and then proposes a solution to address them.

Our work has been impacted by "Deep learning". In the recent years, deep learning has made significant strides due to availability of larger datasets, powerful hardware and the development of advanced training techniques. Ever since Alex Krizhevsky *et al.* [66] won the imagenet image classification challenge [24] in 2012 with a end-to-end trained (feature and classifier) deep convolutional architecture, the field has grown tremendously and deep learning based techniques have been successfully applied to various problems of recognition [135], reconstruction [27, 30] and reorganization [50] (popularly referred as three R's of computer vision [92]).

We have explored both shallow and deep features in this thesis. While our detection work in chapter 3 is based on dense-SIFT features, the rest of our recognition work are based on deep CNN features. For experimentation with classical face datasets such as Yale and AR, raw pixel features are used as per the popular convention.

When demonstrating the issues of limited labeled samples and domain mismatch, we consider person recognition using facial cues, and focus on improving the performance using semi-supervised framework. The approaches however are generic and can be extended to recognition using multi-body cues.

1.5 Thesis Outline

In chapter 2, we briefly review background works that motivated us to develop new solutions to our problems of interest.

In chapter 3, we consider the problem of face detection in-the-wild exemplar face detection. We propose two key improvements for exemplar detectors based on visual word spatial context and co-occurrence that improve voting process. We show the improvements obtained with our proposed approach on standard face detection benchmarks of FDDB [56], AFW [179] and G-album [37] datasets and provide comparisons with state-of-the-art approaches.

In chapter 4, we consider practical issue of availability of training data in the context of person recognition. We demonstrate how unlabeled data can be leveraged in such cases using a semi-supervised learning technique to improve the recognition performance. The experimental results conducted on classical face recognition datasets such as Yale [40], AR [94] and CMU-PIE [125] datasets are shown.

In chapter 5, we consider the issue of domain mismatch between training and test images relevant in practical person recognition. We consider an experimental setup whose objective is to recognize subjects in a video (such as movie tracks) given a training gallery of still images (such as IMDB or publicly available celebrity datasets). We describe our proposed approach that identifies seed images from the target domain and exploits them to identify other target images. The effectiveness of our approach is demonstrated on several challenging datasets.

In chapter 6, we motivate the problem of full body person recognition using multiple body cues when face is not completely visible. Here, we propose a person recognition approach based on deep training and adaptive aggregation of several pose-specific multi-region representations. Along the same direction, we introduce an end-to-end deep person recognition architecture in chapter 7 that extracts features from multiple body regions and then adaptively aggregates them in an end-to-end fashion. We create several person recognition benchmarks and demonstrate the effectiveness of our proposed pose-aware and end-to-end approaches.

Finally in chapter 8, we summarize the thesis, derive conclusions and mention about the directions for future research.

Chapter 2

Background

In this chapter, we give a brief overview of the problem space we are interested in, and describe popular datasets and benchmarks associated with each of these problems. We then describe several successful face/person detection and recognition algorithms. In the last section, we provide an overview of semi-supervised learning techniques that are used in our work.

2.1 Problem Space and Datasets

Face detection: The goal of face detection is to determine the presence of faces in a given image and then return their location in terms of their bounding box co-ordinates as shown in Figure 2.1(a). The task is challenging due to variations such as lighting, pose, orientation, resolution, occlusions, imaging conditions, *etc*, seen in faces. Some of the popular face detection datasets are given below.

- `FDDB` [56] is the most popular face detection benchmark designed for studying the problem of unconstrained face detection. FDDB images are collected from Yahoo news website and contain face annotations for $5{,}171$ instance appearing in $2{,}845$ images. The popularity of the benchmark is also due to its dedicated website that details experimental and evaluational protocol, and access to previously published results, making it easier for researchers to compare different results. One of the main challenges of the FDDB images is the resolution of images. It contains faces as small as 20 pixels in highly challenging and cluttered scenes.
- `AFW` [179] contains $4{,}68$ faces present in 205 images. The database is characterized by cluttered background with pose, aging, and occlusion variations. The dataset provides facial landmarks in addition to facial ground truth boxes.

- WIDER FACE [156] is the latest benchmark created as the current algorithms are achieving near 100% performance on FDDB and AFW. The dataset is much larger consisting of 32,203 images and 393,703 faces. The instances have high degree of variability in scale, pose and occlusion not seen in FDDB. The dataset contains separate training, validation and testing splits, and provides an evaluation protocol similar to FDDB. In the coming years, face detection algorithms will be primarily evaluated on this dataset.

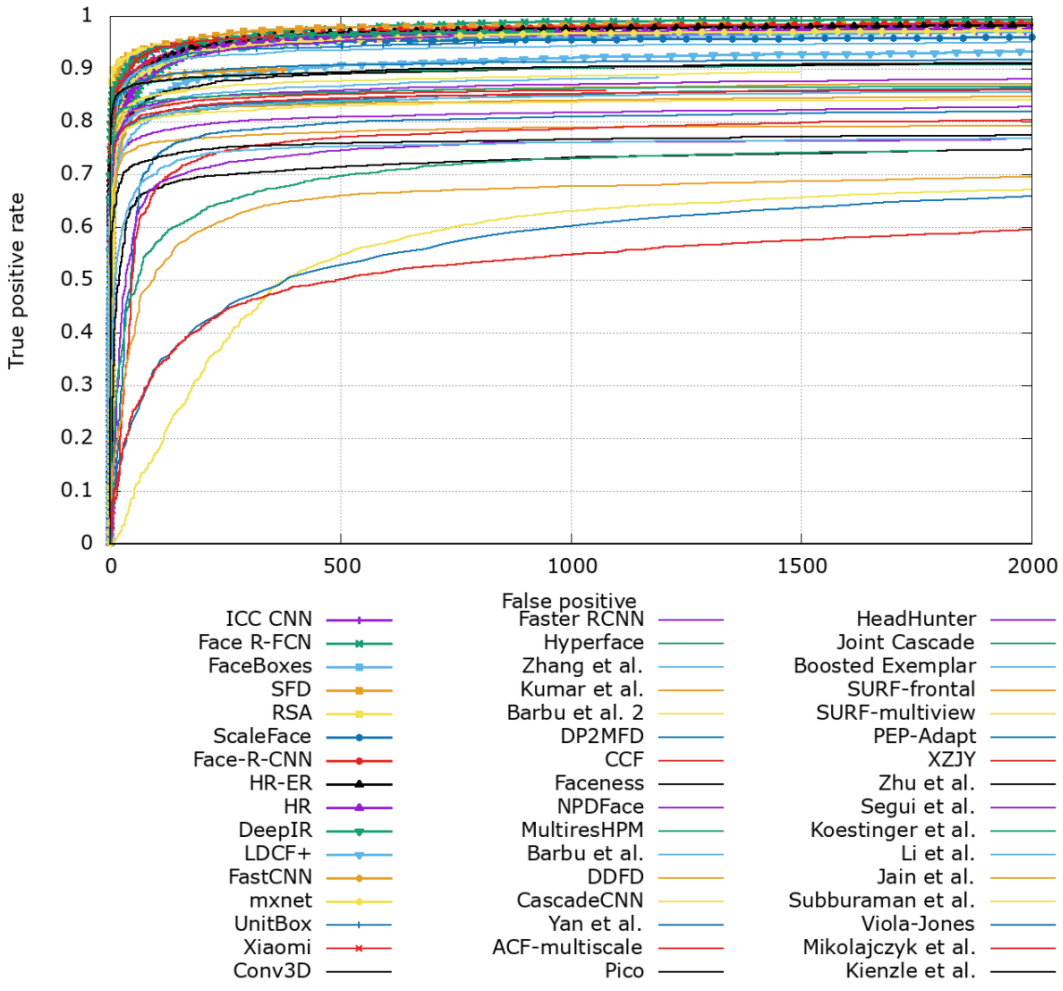
Person Detection: The goal here is to detect complete body region of people. Person detection is very relevant in autonomous driving and surveillance community where it is necessary to detect people even when faces are not completely visible. Below are some of the popular pedestrian detection datasets.

- INRIA [23] is amongst the oldest dataset released by the authors of popular histogram of gradient features. The dataset has comparatively few images with high quality annotations of pedestrians in diverse settings. The dataset is primarily used for training purposes.
- Caltech-USA [26] is a larger pedestrian detection dataset consists of several videos captured from real traffic scene. It consists of approximately 10 hours of video clips taken from a vehicle driving through regular traffic in an urban environment. There are about 250,000 frames with a total of 350,000 bounding boxes and 2,300 unique pedestrians were annotated.

Face Recognition: The objective here is to predict the identity of a person from his/her cropped face. Unlike in detection setting, recognition systems need training (gallery) examples of the subjects that need to be recognized during testing. Below are the popular facial recognition datasets.

- Classic datasets: Yale [40], AR [94] and CMU-PIE [125] are the oldest face recognition datasets that were used extensively until recently. These datasets are created in constrained lab settings where images are captured under pre-defined illumination, pose, and expression variations. Due to their smaller size and constrained variations, these datasets are not popular in the recent years due to the emergence of deep learning techniques that require large training data.
- IMFDB or Indian Movie Face database [119], created by us, is a large unconstrained face database consisting of 34,512 images of 100 Indian actors collected from more than 100 videos. The images are manually selected and cropped from the video frames resulting in a high degree of variability in-terms of scale, pose, expression, illumination, age, resolution, occlusion, and makeup. The database provides detailed annotation in terms of face bounding box, and meta data such as

(a)

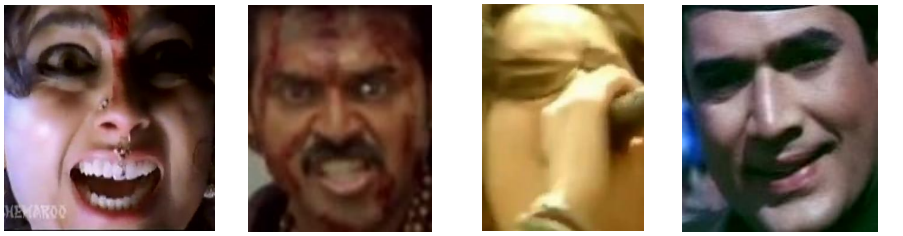


(b)

Figure 2.1 (a) Few examples from the face detection dataset and benchmark (FDDB) dataset. (b) Performance of various face detection algorithms measured on FDDB dataset [56]. More details of these results can be found at `http://vis-www.cs.umass.edu/fddb/results.html`

(i)

(ii)



(iii)

Figure 2.2 Indian Movie Face Database (IMFDB). (i) Images of actor *Amitabh Bachchan*. Each actor in IMFDB database contain images with diverse variations in age, pose, expressions, resolution, illumination. (ii) shows appearance variations based on (a) age, (b) make-up, (c) illumination, (d) pose, (e) expressions and (f) Occlusions. (iii) Examples showing detailed annotation with meta-data information for each image.

age, pose, gender, expression and type of occlusion. Few images from the database and annotations are shown in Figure 2.2

- `Mega Face` [98] is the latest database created to facilitate training of deep convolutional neural networks for person identification. The dataset is created from Flickr images and consists of 4.7 million instances belonging to 672K subjects. In terms of scale and size, it is the largest publicly available database today.
- `Movie Trailer` [101] is created for studying face recognition in videos. The dataset consists of 101 movie trailer videos released in the year 2010 and contain $4{,}485$ face tracks. The subjects in this dataset overlap with another face recognition dataset - `PubFigs` [67]. A popular experimental procedure followed when evaluating on videos uses movie trailer as test set with `PubFigs` as gallery set.

Face Verification is another topic related to face recognition. Given a pair of images, the task is to predict whether the two images belong to the same person or not. The problem has received great attention in the recent years due to the creation of verification benchmarks.

- `LFW` or Labeled faces in the wild [55] contains $13{,}233$ facial images of $5{,}749$ people collected from the web. The benchmark consists of training, validation and testing splits each containing mutually exclusive image pairs. Like `FDDB`, `LFW` has become the popular choice for face verification due to dedicated website for submitting results and comparisons.
- `PubFigs` [67] is created in similar spirit to `LFW` and contains $58{,}797$ images of 200 people. While `LFW` contains large number of subjects and fewer images per subject, `PubFigs` contains more instances for each subject.
- `YouTube Faces` [148] is designed for studying face verification in videos. It contains $3{,}425$ videos of $1{,}595$ different people downloaded from YouTube. It follows experimental setup and evaluation protocol similar to `LFW`.

Person Re-identification is the task of matching pedestrians captured in non-overlapping camera views, a primary requirement in surveillance. The task has primarily emerged in the recent years due to increased necessity for public safety and widespread deployment of camera networks in public places such as parks, universities, streets, *etc*. The most popular datasets in this space are `VIPeR` [46], `CAVIAR4REID` [20] and `CUHK03` [79].

Person Recognition methods are proposed for identifying people in unconstrained settings where faces are not completely visible or in back view. The problem is getting increased attention lately due to the introduction of PIPA [165] dataset. The dataset consists of very challenging examples and provides head bounding boxes even when the faces are not visible. As a result, algorithms have to rely on multiple body regions to recognize people.

2.2 Face Detection

The area of face detection has made significant progress in the last two decades as can be witnessed in Figure 2.1. We start with a brief description of a few works that were proposed before the seminal work of Viola and Jones [141]. We then describe Viola-Jones framework and other works that followed it. For convenience, we categorize these approaches into four categories namely AdaBoost/decision forests, Deformable parts model, Exemplar and Deep network based approaches. A brief comparison of these approaches is given in Table 2.1 and Table 2.2.

2.2.1 Pre Viola-Jones Detectors

Sakia *et al*. [115] made one the first attempts[1] to detect frontal faces in photographs in the year 1969. They modeled the face by creating templates for different face regions such as eyes, nose, mouth, and contours. Each template is defined in terms of line segments. The contour template is applied on the input image to get the probable locations of a face that are further examined using part templates. Since then hundreds of papers have been proposed for face detection, which fall into four categories namely knowledge-based methods, feature invariant approaches, template matching methods, and appearance-based methods. A complete review of these methods is provided in [155]. Knowledge-based methods use pre-defined rules to determine a face based on human knowledge. Feature invariant approaches aim to find facial features that are robust to pose and lighting variations. Features such as edges, skin color, and texture or a combination of these features are used to perform detection. While template matching methods use hand-crafted face templates to detect faces, appearance based methods follows data driven approach and learn face models from training images to perform detection. Appearance based methods perform superior to all the other approaches owing to availability of large datasets and increased computational capacity.

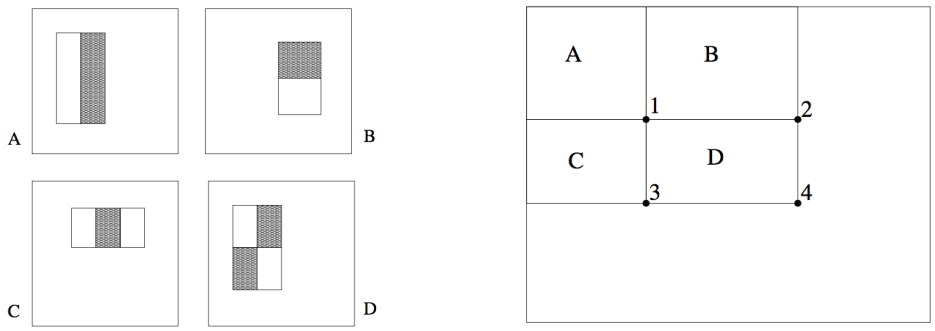
[1] as per Google Scholar

Table 2.1 A brief summary of various face detection approaches measured in terms of recall@2000 false positives on FDDB [56] benchmark.

Family	Method	Description	Features	data	train time	loss	Recall
AdaBoost	Viola-Jones (opencv implementation) [84]	Real-time system consisting of cascade of adaboost classifiers trained on binary Haar features. Fast inference due to integral image technique for feature computation and cascade structure for rejection of background regions.	Haar	5000	$2-5$ days	false positive rate	65.92% (frontal model)
AdaBoost	PICO [93]	Identical to VJ but the focus is on deployment in low computer power such as mobiles. Tests at each node involve simple pixel intensity comparison as opposed to Haar feature response comparison.	pixel intensity	20000	$2-3$ days	false positive rate	74.68%
AdaBoost	NPD [83]	Features based on normalized pixel difference. The features can be efficiently obtained from a look up table resulting in faster runtime.	NPD	26000	–	tree depth as parameter	82.78%
AdaBoost	Surf Cascade [78]	Improvises VJ with the replacement of Haar with SURF features, objective function from false PR to area under the curve (AUC), and the use of large training data and fewer cascade stages.	SURF	13000	<1 hour	AUC	84.08%
Decision forests	Aggregate Channel Features [153]	Strong features based on multiple channels such as color channels (gray, RGB, HSV, LUV), gradient magnitudes and histograms, weak classifiers based on depth-2 decision trees and a single stage cascade are used in VJ framework	color channels, gradient magnitude & histogram	36112	10 mins	number of weak classifiers as training parameter	86.07%
Decision forests	Integral Channel Features [96]	similar to [153] with slightly different set of channel features.	RGB, RmG, L, U, gradient and skin	26000	<1 hour	number of weak classifiers as training parameter	85.7%
DPM	Supervised DPM [179]	Performs simultaneous face detection, landmark and pose estimation. Consists of mixtures of trees with a shared pool of parts. A part is defined on each facial landmark and topological changes in viewpoint are captured using global mixtures.	HOG	900	$2-4$ hours	structured SVM	77.43%
DPM	HeadHunter [96]	General object detector DPM [32] trained on larger face dataset with NMS=0.3	HOG	26000	$2-8$ hours	latent SVM	87.1%

Table 2.2 `Continued`. A brief summary of various face detection approaches measured in terms of recall@2000 false positives on `FDDB` [56] benchmark.

Family	Method	Description	Features	data	train time	loss	Recall
DPM	Fastest DPM [151]	DPM made faster by imposing rank constraints on root filter, cascade based pruning of low hypothesis regions and look-up table for HOG feature computation	HOG	26000	$2-8$ hours	latent SVM	86.15%
DPM	DP2MFD [108]	DPM with deep convolutional features	CNN	2500	$2-4$ hours	linear SVM	91.2%
Exemplar	Exemplar [76]	Detection using retrieval framework. Voting maps generated from a large pool of indexed training exemplars are aggregated to locate faces.	SIFT	18486	$8-12$ hours	–	80.25%
Exemplar	Boosted Exemplar [123]	Similar to [76] but uses a small set of 'discriminative' exemplars selected selected through adaboost training. Improved detection procedure with tile-based detection and feature re-usage at different scales.	SIFT	3000	$8-12$ hours	–	85.65%
Deep Networks	Cascade CNN [77]	Cascade of three small CNN networks, each network tuned to reject low confident regions. These networks process images at 12, 24 and 48 window size, respectively. Each network is followed by a calibration network that adjusts the bounding boxes.	CNN	26000	–	softmax	85.67%
Deep Networks	Hyper Face [109]	Multi-task CNN. A single CNN network is trained to perform joint face detection, landmark, pose and age estimation.	CNN	26000	–	sum of two Euclidean & softmax losses.	90.86%
Deep Networks	Conv3D [82]	Generates face proposals by estimating facial keypoints and a fixed 3D face template. A bounding box regressor is applied on these proposals to locate the faces.	CNN	26000	–	softmax + l_1 loss	91.18%
Deep Networks	Faster RCNN [60]	Faster R-CNN trained for face detection. Consists of object proposal and classification layers.	CNN	159424	–	softmax + l_1 loss	96.11%
Deep Networks	Tiny Faces [54]	Targets face detection at small scales. Uses coarse image pyramid (0.5, 1x, 2x) and 25 templates to detect faces of different sizes. Each template may have different resolution capturing different amount of context. Templates for smaller faces use large amount of context to improve performance.	CNN	159424	–	softmax + l_1 loss	98.31%



(a) (b)

Figure 2.3 (a) **Haar-features:** The sum of the pixels in the white rectangles are subtracted from the sum of pixels in the grey rectangles. Features can be 2- (A and B), 3- (C), or 4 (D) rectangular. (b) **Integral Image:** The value at any location of the integral image is the sum of pixels *above* and *left* to it. The value at D is (A+B+C+D). The sum of pixels within D can be computed as $(4 + 1) - (2 + 3)$. Images are taken from [141].

2.2.2 AdaBoost Detectors

Viola and Jones proposed a breakthrough work based on boosting for face detection in the year 2001. Unlike its contemporary works, the approach is data-driven and runs in *real-time*, thus paving the way for deployment in real-world applications such as digital cameras. The paper introduces three innovative ideas - (a) *integral image* representation that allows the features to be computed efficiently (b) Adaboost training algorithm with binary Haar-like features, and (c) *Cascade-structure* to quickly discard background regions and focus on object-like regions.

The features are based on three simple binary features (shown in Figure 2.3(a)) that can be efficiently computed using an intermediate representation called "integral image". The integral image at location (x, y) contains the sum of pixels above and to the of left of (x, y) i.e. $ii(x, y) = \sum_{\bar{x} \leqslant x, \bar{y} \leqslant y} i(\bar{x}, \bar{y})$ where $ii(x, y)$ is the integral image of the original image $i(x, y)$. Once the integral image is computed, any rectangular sum can be calculated in constant time as shown in Figure 2.3(b).

Adaboost training procedure involves optimizing the true positive and false positive rates by iteratively selecting "discriminative" features from a large pool of candidate features. A cascade structure is employed to make the detection process faster, wherein the background regions are evaluated and rejected in the initial stages. The latest and close variant of VJ is the SURF cascade [78] that use SURF features and area under the curve as objective function for training the cascade.

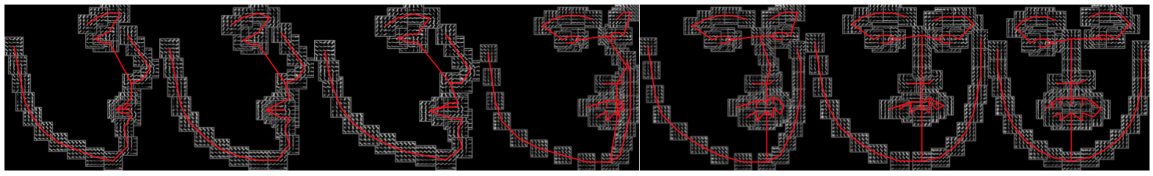


Figure 2.4 Deformable parts model for face detection. Each face is composed of several parts (landmarks) with a springs between them. Large number of parts are learned for multiple views and then each tree is allowed to select parts from a common, shared pool of parts that makes learning and inference efficient. Figure is from [179].

Integral channel features [96, 153], which are another variants of VJ have achieved high performance recently in the challenging conditions. These approaches make following modifications to the VJ framework. First, instead of using single grayscale channel as input, multiple feature channels consisting of RGB, HSV, LUV color channels, gradient magnitudes and histograms are used. Second, weak classifiers are changed from decision stump to depth-2 decision tree to achieve better discriminative ability. Third, a single stage of Adaboost classifier rather than a cascade structure to achieve faster detection. These changes result in faster training and detection with improved performance.

2.2.3 Parts Based Model

Deformable parts models (DPM) [32, 34, 48, 103] is another popular line of work that provides an elegant framework for detecting faces and humans with high intra-class variations. DPMs treat objects as being composed of several deformable parts, and thus model the appearance of individual parts along with their spatial configurations. For e.g., DPM models the appearance of eyes, nose, mouth, ear for human face along their spatial positions relative to each other.

In the DPM framework [32], the model for an object with n parts is defined using a root filter F_0 and a set of part models $(P_1, \ldots, P_n)$ where $P_i = (F_i, v_i, s_i, a_i, b_i)$. Here F_i is a filter for the i-th part, v_i is a two-dimensional vector specifying the center for a box of possible positions for part i relative to the root position, s_i gives the size of this box, while a_i and b_i are two-dimensional vectors specifying coefficients of a quadratic function measuring a score for each possible placement of the i-th part. The score of a placement is given by the scores of each filter (the data term) plus a score of the placement of each part relative to the root (the spatial term),

$$\sum_{i=0}^{n} F_i \cdot \phi(H, p_i) + \sum_{i=1}^{n} a_i \cdot (\bar{x}_i, \bar{y}_i) + b_i \cdot (\bar{x}_i{}^2, \bar{y}_i{}^2),$$

where $(\bar{x_i}, \bar{y_i}) = ((x_i, y_i) - 2(x, y) + v_i)/s_i)$ gives the location of the i-th part relative to the root location. The filters (F_i) and parameters (a_i, b_i, v_i) are learned during training using Latent SVM framework that optimizes the objective function in a two-step iterative fashion keeping either the appearance of parts or their positions fixed.

Although DPM was originally proposed for person detection, Zhu and Ramanan [179] extended it for face detection. Their model (as shown in Figure 2.4) consists of a mixture of trees and a pool of parts, with parts being shared across the trees. Unlike vanilla DPM that learns the parts and their positions, parts are modeled using extra-supervision of facial landmarks. To allow flexibility and reduce the redundant parts during training, mixtures are allowed to share the parts. The pool consists of large number of parts covering all possible variations of face regions across landmarks. The framework models complex face variations across viewpoints to detect faces in challenging scenarios. Very recently `HeadHunter` [96] showed that vanilla DPM without any landmark supervision can achieve superior results over [179] with landmarks, when properly trained on a large training set.

2.2.4 Deep Networks for Detection

Cascade CNN: Inspired by the success of convolutional networks in the recent years, Li *et al.* [77] proposed the CNN based architecture for face detection. Their system consists of three smaller CNN networks named $12-net$, $24-net$ and $48-net$ as shown in the Figure 2.5. They follow a cascade style similar to VJ, wherein each net processes the outputs of the previous net. Given a test image, the $12-net$ first scans the whole image densely across different scales and quickly rejects most of the background regions. The remaining candidate detection windows are cropped out and resized into 24×24 as input images and fed to $24-net$ that processes it and rejects any non-face regions. The last $48-net$ accepts the passed detection windows as 48×48 images and evaluates each of these detection windows. In addition, the system consists of 3 similar calibration networks that improve the localization of the bounding boxes obtained at each stage.

Faster R-CNN: The state-of-the-art CNN visual object detection system is "region-based convolutional neural networks" or popularly known as R-CNN [43]. The architecture of R-CNN is shown in Figure 2.6. Instead of scanning the images for desired objects through a sliding window, it gathers a large set of category-independent region proposals and then applies a deep classifier trained for classification task. The approach though simplistic is very powerful and its latest variants are currently the state-of-the-art for object detection. The proposals can be generated either using off-the-shelf pro-

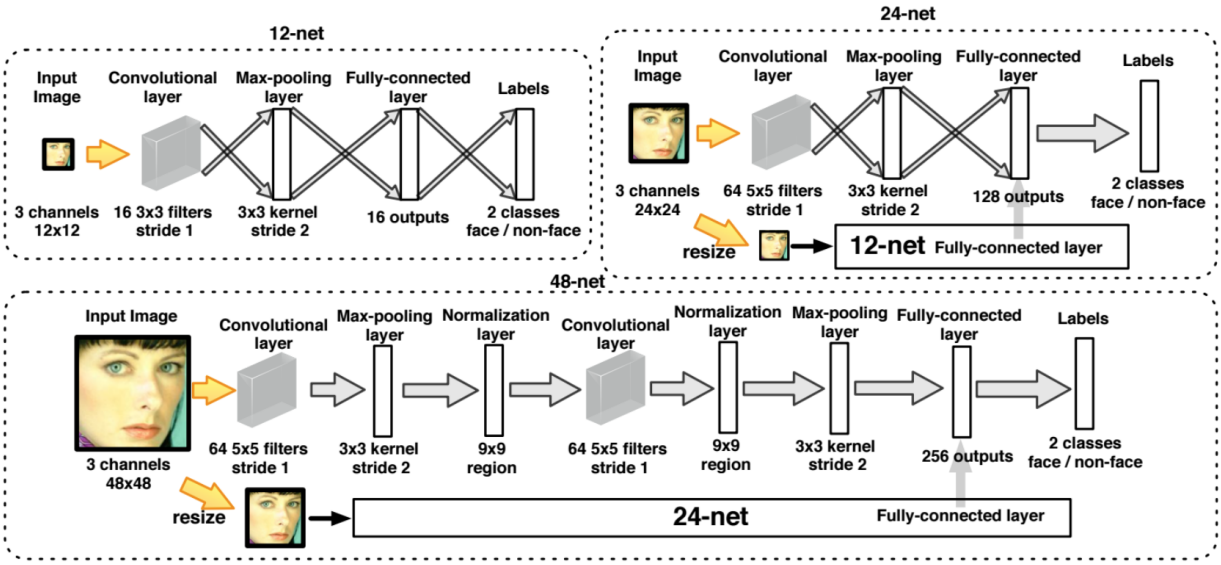


Figure 2.5 Cascade CNN: A Convolutional Neural Network Cascade for Face Detection. The system consists of cascade of three convolutional networks processing inputs at 12, 24 and 48 pixels. Each network processes the input regions and forwards only the high confident regions to next network for further processing. Figure is from [77].

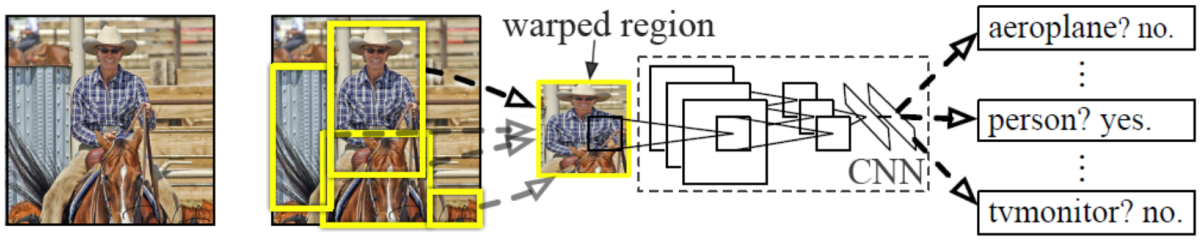


Figure 2.6 R-CNN framework for object detection. It consists of two components namely object proposal generation and classification of proposals using a CNN trained for multi-class classification. The proposals can be obtained using off-the-shelf object proposal algorithms or obtained through end-to-end training in a single network. Figure is from [43].

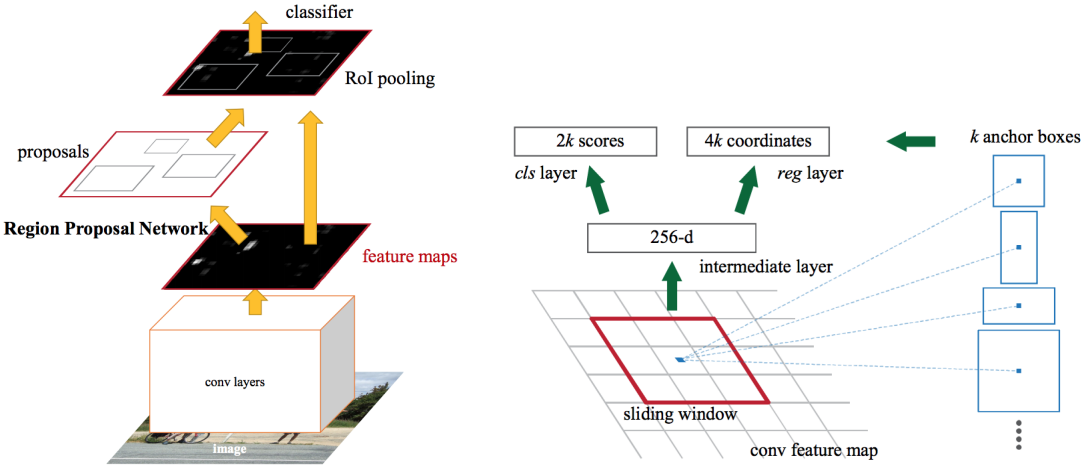


Figure 2.7 Faster RCNN. (a) End-to-end detection architecture with proposal generation and classification. (b) Region proposal network outputs at each location of feature map utmost k object proposals relative to k fixed anchor boxes of multiple sizes and scales. For each proposal, class probability scores and its relative locations to anchor boxes are predicted. Figures are taken from [112].

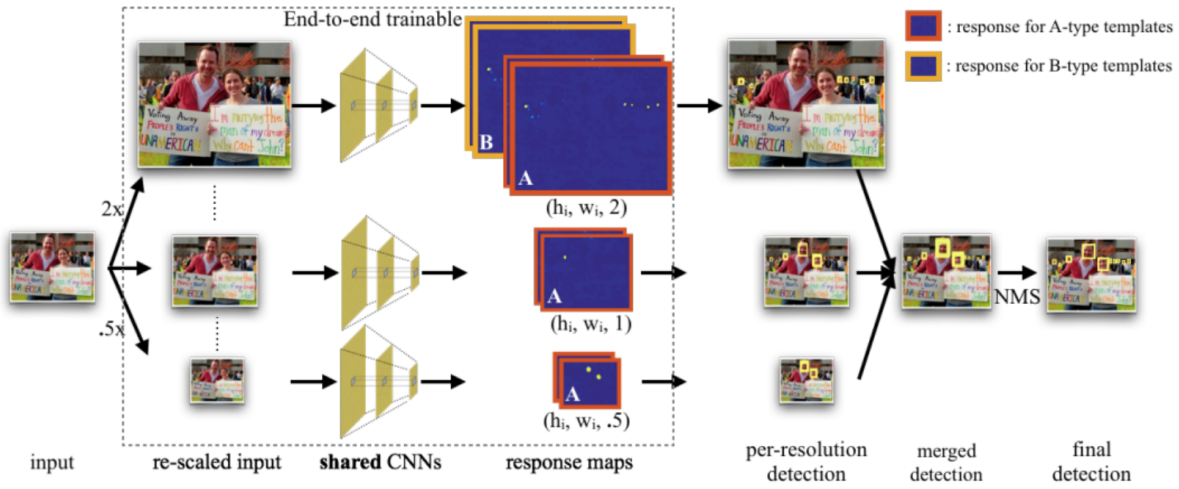


Figure 2.8 Tiny faces architecture for face detection [54]. A coarse pyramid of images at three scales: $0.5X$, X, $2X$ are fed into the network to predict template responses for both detection and regression at every resolution. Large number of templates corresponding to different scales are applied from the feature maps at different scales. It is observed that by considering large amount of context significantly improves the performance for detecting smaller faces. Figure is from [54].

posal generators such as selective search [138] or can be generated using a proposal generation network through an end-to-end detection algorithm known as Faster R-CNN [112].

Faster R-CNN network [112] consists of a series of convolutional layers shared across region proposal generation network (RPN) and a classification network. An RPN is a fully convolutional network that simultaneously predicts object bounding boxes and their class scores at each position as shown in Figure 2.7. The RPN is trained end-to-end to generate k region proposals at each location with respect to k pre-defined reference anchor boxes of different size and scale. Huaizu and Miller [60] recently applied faster R-CNN for face detection replacing the final classification layer to perform two-class classification problem (face/no face). With this simple modification, they achieved impressive performance when trained on large dataset `WIDER FACE` [156].

Tiny Faces: While most approaches detect faces at large scales with high degree of accuracy, the gap becomes wider when detecting faces at extremely small scales. Most of the approaches treat different sized faces similarly and adapt a common strategy to detect them all. However, it is shown that such a strategy may not work well to detect faces of different sizes as the facial cues are fundamentally different at different scales. In [54], an end-to-end approach that can detect faces at different scales is proposed. It uses a coarse image pyramid at three scales (0.5, 1x, 2x) and 25 templates to detect faces of different sizes. Each template defines a fixed window with different resolution to capture context. Templates for smaller faces use large amount of context to improve performance. It is also shown that features have to be extracted from multiple layers of deep networks to detect faces at different scales.

2.2.5 Exemplar Based Approach

Exemplar based detectors [76, 123, 69] follow retrieval framework for detection. In this paradigm, each exemplar generates a voting map that are combined to locate the faces in the target image as shown in Figure 2.9. During the training phase, a large database of exemplars that cover significant face variations are collected. Local features (such as SIFT) are extracted and k-means based vocabulary is constructed followed by feature quantization. Term frequencies (TF) and inverse document frequencies (IDF) are calculated and inverted files are created similar to BoW retrieval scheme. During testing, each exemplar generates voting maps at multiple scales using the spatial locations of the features. Each location in the map indicates the similarity score between the exemplar and the image sub-region at that location. Given a exemplar e_i and the rectangular region centered at location p of the test image x, the

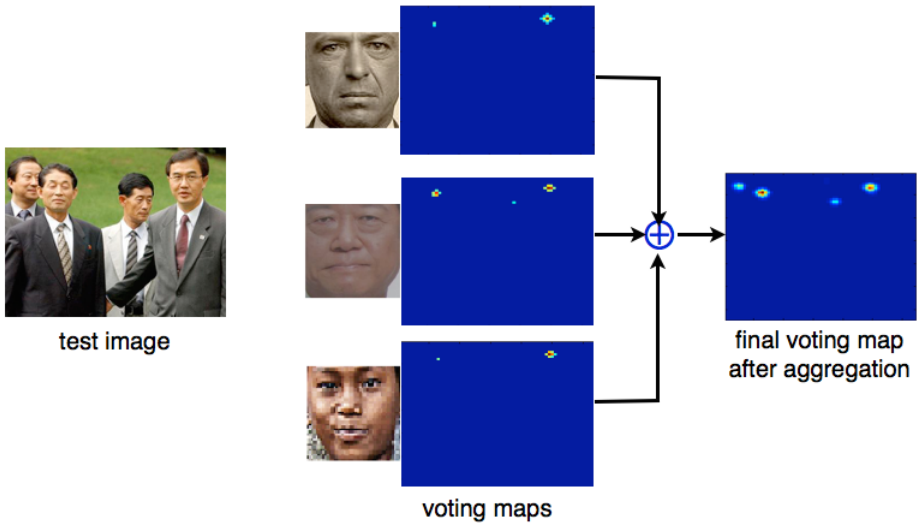


Figure 2.9 Ensemble of Exemplars for Face detection: A large database of diverse exemplars is collected and indexed using a BoW representation. During testing, each exemplar casts a vote on the test image at multiple scales. The votes from different exemplars are then aggregated to detect the faces.

similarity measure between them is given as

$$S(p, e_i) = \sum_k \sum_{\substack{f \in R_x(p), g \in e_i \\ w(f)=w(g)=k \\ ||\mathbb{T}(L(f))-L(g)||<\epsilon}} \frac{idf^2(k)}{tf_{e_i}(k) \cdot tf_x(k)},$$

where x is the test image, $R_x(p)$ is the sub-image region of x centered at p. f and g are the local features and $L(f)$ and $L(g)$ are their corresponding locations from x and e_i, respectively. $w(f)$ and $w(g)$ are the quantized visual words of features f and g respectively. $w(f) = w(g) = k$ indicates that only the matched visual words are considered for voting. The spatial constraint $||\mathbb{T}(L(f)) - L(g)|| < \epsilon$ further ensures that matched features should be closer under some unknown transformation $\mathbb{T}$. To handle burstiness, weights are divided by $tf_{e_i}(k)$ and $tf_x(k)$, which denote the TF of the visual word k in the exemplar and test image, respectively.

Finally, gating is applied on each of the voting maps by subtracting the scores with a pre-trained threshold. The resulting maps from different exemplars are then aggregated to obtain the final voting map. Since each exemplar is *specific* to particular variation, it is possible to detect faces in challenging conditions using a sufficiently large exemplar database. The advantage of this approach is that by having a large database that covers all possible variations, faces in challenging conditions can be detected without having to learn explicit models for different variations. The approach is scalable and offers

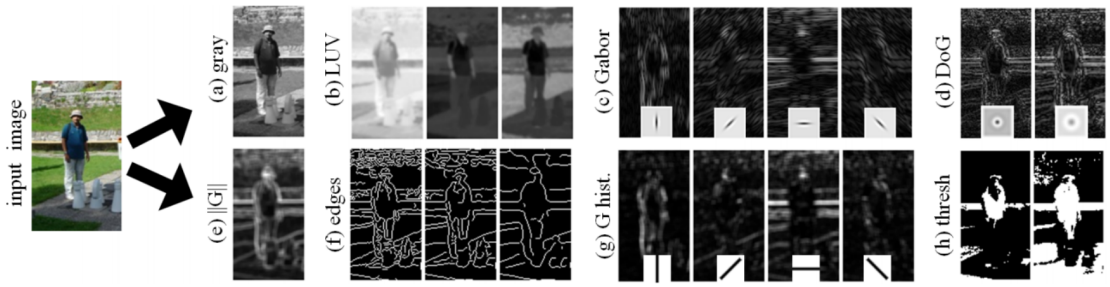


Figure 2.10 Integral channel features in combination with decision forests work well for pedestrian detection. Figure is from [96].

flexibility to add more exemplars without additional training required. It is also efficient as it avoids the exhaustive sliding-window search typically employed in model-based approaches.

Summary: Face detection has witnessed tremendous progress in the last decade. While the cascade Adaboost classifiers with Haar features paved the way for real-time deployment, its subsequent variations with discriminative features such as SURF provided improvement in detection performance. On the other hand, exemplar and DPM models perform well in the challenging in-the-wild scenarios due to their sophisticated detection framework consisting of bag of exemplars or part detectors. Deep convolutional networks are currently the best performing methods, and outperform all the previous approaches that use shallow features.

2.3 Person Detection

Person detection has received similar attention in last two decades because of its direct applications in car safety, surveillance, and robotics. The methods that worked well successfully for face detection are applied for person detection and vice-versa. Similar to face detection, the three major directions in person detection are based on (a) decision forests (b) DPM variants and (c) deep networks. These algorithms were covered in detail in the previous section. Rodrigo *et al.* [11] make a thorough comparison of these approaches and make following conclusions. As shown in Figure 2.11, decision forest based algorithms perform the best for person detection although it is not clear what gives them the extra edge. The decision forests seem to work best with integral channel features as shown in Figure 2.10.

The latest trend in person detection or object detection in general is to avoid time-consuming sliding window and develop an end-to-end deep neural network that are based on proposal generation. R-CNN detectors [43, 112] described in the previous section are one family of detectors with superior detection

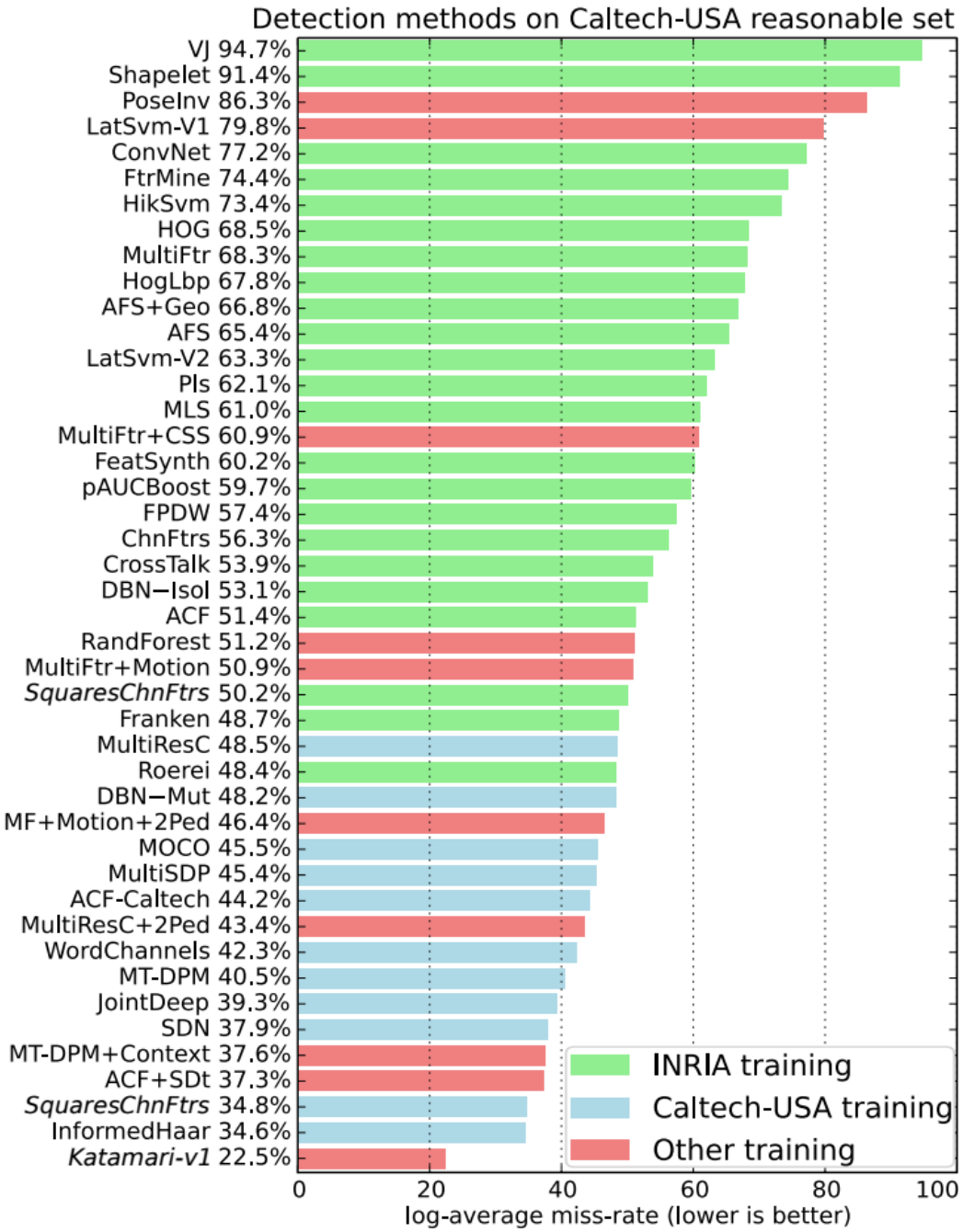
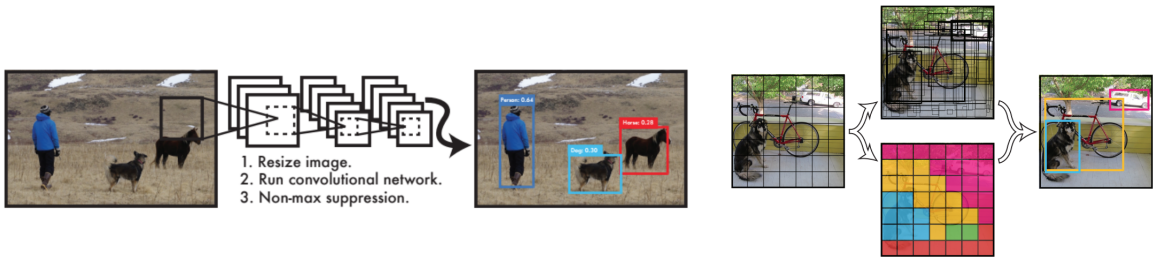


Figure 2.11 Person detection over last 10 years on Caltech-USA pedestrian dataset. Decision forest based algorithms with integral channel features perform best for person detection. Figure is from [11].



(a) (b)

Figure 2.12 (a) YOLO detection system and (b) YOLO formulates detection as a regression problem. It divides the image into an even grid and simultaneously predicts bounding boxes, confidence in those boxes, and class probabilities. Figure is from [110].

and run-time performance. We below describe two other state-of-the-art object detectors based on deep neural networks that produce impressive result on person detection task.

YOLO: You Only Look Once (YOLO) [110] is the state-of-the-art end-to-end neural network that detects multiple objects (including people) appearing in an image at extremely faster rate. Contrary to sliding window or proposal based approaches that repurpose classification networks for detection tasks, YOLO treats detection as a regression problem and generates the bounding boxes of different objects along with their class probabilities in one single evaluation (Figure 2.12).

As shown in Figure 2.12, YOLO divides the input image into a square grid ($S \times S$) wherein each cell predicts B bounding boxes defined as (x, y, w, h, s) along with the confidence scores of C objects. This is achieved using a fully connected layer that outputs fixed size tensor ($S \times S \times (5 \times B + C)$). Due to its unified framework without multiple steps involved for proposal generation, feature pooling or bounding box refinement, YOLO achieves impressive speed of 45 FPS at 63.4 mAP on Pascal VOC 2007. A new version named YOLOv2 [111] released recently further improves the performance to 78.6 mAP and runs at 40 FPS by incorporating strategies such as batch normalization, priors on anchor boxes, multi-scale training and simpler network architecture.

Single Shot MultiBox Detector (SSD): SSD [87] is another state-of-the-art detector that is very similar in design to YOLOv2. Similar to YOLOv2, it generates a fixed size output grid wherein each cell predicts the confidence of objects and their bounding boxes with respect to fixed number of anchor points. Contrary to YOLOv2, it produces predictions of different scales from feature maps of different scales. It achieves this by using multiple convolutional layers at the end of a base network (such as

VGG) that generate feature maps that decrease in size progressively. On Pascal VOC 2007, SSD achieves 74.3 mAP with 300×300 input at 59 FPS and 76.9 mAP on 512×512 input at 19 FPS.

2.4 Face Recognition

Face recognition that aims at recognizing identities of the subjects is another widely studied topic in computer vision due of its applications in access control, security, surveillance, photo tagging, image search, etc. The earliest works proposed for automatic face recognition that dates back to 60s and 70s are [14, 63]. They focus on detecting individual features such as the eyes, nose, mouth, and head outline, and defining a face model by the position, size, and relationships among these features. Since then, thousands of papers have been proposed for face recognition. A survey [171] done in 2003 broadly groups different face recognition approaches into three different categories- holistic methods, feature based approaches and hybrid methods. Holistic methods use the entire face region for recognition. Feature-based methods extract local features around discriminative regions such eyes, nose, and mouth before feeding to a classifier. Hybrids methods makes use of both local features and whole face image to predict the identities. We next give a brief overview of works that are influential, impactful and closely related to our work.

2.4.1 Eigen Faces

Eigen Faces [137] introduced by Turk and Pentland in 1991 is the first really successful demonstration of automatic recognition of faces. The main intuition behind the approach comes from the information theory that deals with encoding the given data to obtain a compact and efficient representation. Turk and Pentland wanted to extract only the "relevant" information from faces and "encode" them as efficiently as possible. Once the compact codes are obtained for every face, a simple comparison of them in the new space using some similarity metric can give possible identity of the faces. The compact way of encoding is straightforward - find the directions of the variations in a collection of faces and encode in terms of a few directions of larger variation. The authors use principal component analysis (or Karhunen-Loeve expansion) to find the vectors that best account for the distribution of face images within the entire image space. More specifically, sample vectors ${x_i}^2 \in \mathbb{R}^n$ can be expressed as linear combinations of the orthogonal basis ϕ, $x_i = \sum_j^n a_i^{(j)} \phi^{(j)} \approx \sum_j^m a_i^{(j)} \phi^{(j)}$ (where m$\ll$n) where a_i are

[2] assuming that the face are mean centered and in vector format.

Figure 2.13 Eigen Face: (left) shows images from Yale database and (right) shows top 30 eigen vectors computed on entire Yale database

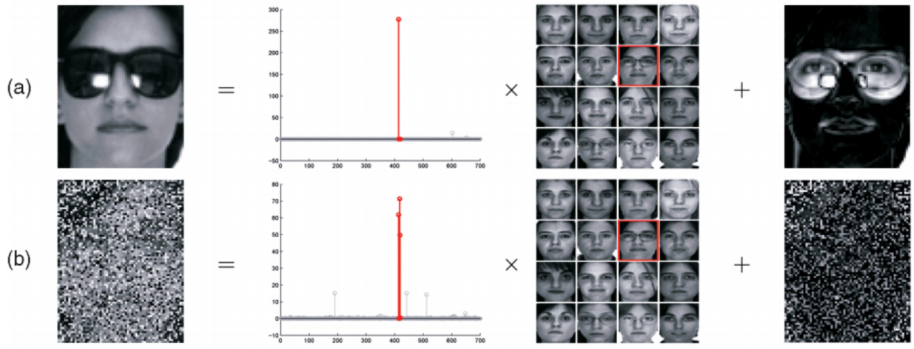


Figure 2.14 Sparse Representation based Classifier represents each test image (left), which is (a) potentially occluded or (b) corrupted, as a sparse linear combination of all the training images (middle) plus sparse errors (right) due to occlusion or corruption. Coefficients in red correspond to training images of the correct individual. Figure is from [150].

the linear weights of face x_i. The basis ϕ is obtained by solving the eigenproblem $C\phi = \phi\lambda$ where C is the covariance matrix of the input sample matrix $X = [x_1, x_2, \ldots, x_N]$.

Figure 2.13(b) shows the obtained eigen vectors obtained for Yale face dataset (Figure 2.13(a)) [72]. Since the obtained eigen vectors have face-like appearance, they are popularly referred as "eigenfaces". Given the eigenfaces, every face in the database can be represented as a vector of weights a_i; the weights are obtained by projecting the image into eigenface components by a simple inner product operation. During testing, the weights of the test image are also obtained. The identification of the test image is done by locating the image in the database whose weights are the closest to the weights of the test image. EigenFaces demonstrated great success in recognizing faces even in the presence of noise such as blurring, partial occlusion and changes in background.

2.4.2 Sparse Representation based Classifier

After the success of sparse representation and compressed sensing for solving various inverse problems in signal processing, Wright *et al.* proposed **Sparse representation based classifier (SRC)** [150] for face recognition. There is a notion of sparsity inherently in the problem since one is interested in finding sample(s) of the single subject out of a large database of subjects that best explains the given query image.

Face images with different illuminations (but in a particular pose) usually lie in a low-dimensional subspace compared to their ambient dimension [40]. Thus if we have a few basis elements that span this low-dimensional subspace, faces of any illumination can be constructed through a linear combination. However, it is not always possible to find these low-dimensional basis. It is reported in [40] that if we have enough training samples with sufficient diversity in illuminations, they lie close to the low-dimensional subspace. Thus, any test image of a subject can be well represented as a linear combination of training images of the same subject. i.e $y = A_i x_i$, where A_i are the images of subject i. Since we will not know the label i in advance during testing, one can still form a linear representation as $y = [A_1, A_2, ..., A_c]x$, where $x = [\dots, 0^T, x_i^T, 0^T, \dots]$

We can see that x is a highly sparse vector with nonzero elements concentrated in images of a single correct subject. Thus face recognition problem reduces to seeking sparsest solution of $y = Ax$. This is illustrated in Figure 2.14. From the compressed sensing literature, one can recover a sparse solution through an l_1 norm minimization reducing the problem to:

$$\hat{x} = \arg\min ||x||_1 \text{ subject to } ||y - Ax||_2$$

Once the sparse coefficients $\hat{x}$ are recovered, test image can be given the label that has largest concentration of non-zero elements or the class that gives minimum reconstruction error. The above method can be made to deal with occlusion and corruption with a minor change.

$$\hat{x} = \arg\min ||x||_1 + ||e||_1 \text{ subject to } y = Ax + e,$$

where e is the corruption or occlusion (like sunglasses) that is mostly sparse.

This sparse representation based algorithm achieved state-of-the-art results on variety of face databases in a constrained setting i.e., when the faces are perfectly aligned. The algorithm produced superior results even in the presence of extreme illumination variations, corruption and occlusion.

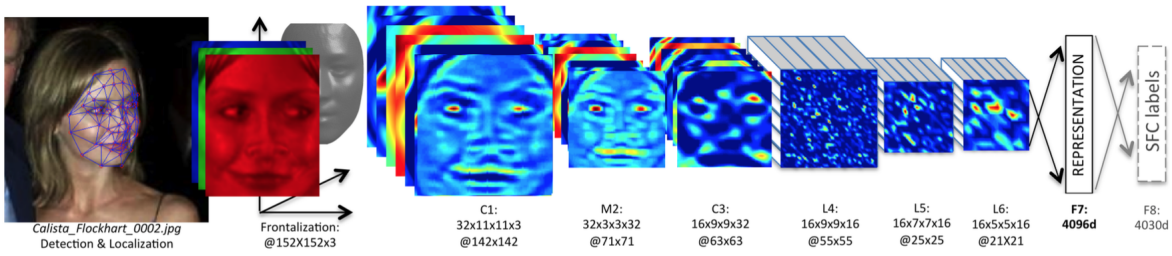


Figure 2.15 Deep Face architecture: It consists of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. The colour maps shows the feature maps produced at each layer. Figure is from [135].

2.4.3 Deep Networks for Recognition

One of the earliest CNN based face recognition papers is [71], which appeared in 1997. After more than 15 years, *DeepFace* [135] developed by Facebook with almost a similar architecture to that of [71] achieved human-level accuracies on LFW face verification[3] benchmark. The primary difference being the use of massive amounts of training data and efficient training strategies.

Figure 2.15 shows DeepFace CNN architecture [135]. The 152×152 faces after alignment are fed into CNN network. The first layer is a convolutional layer (C1) consisting of 32 filters of size $11 \times 11 \times 3$. The resulting 32 feature maps are then fed to a max-pooling layer (M2) that takes the max over 3×3 spatial neighborhoods with a stride of 2, separately for each channel. This is followed by another convolutional layer (C3) that has 16 filters of size $9 \times 9 \times 16$. Max-pooling layers make the output of convolution networks more robust to local translations. The subsequent layers (L4, L5 and L6) are instead locally connected, like a convolutional layer they apply a filter bank, but every location in the feature map learns a different set of filters. The last two layers (F7 and F8) are fully connected. The output of the last fully-connected layer is fed to a K-way soft-max (where K is the number of classes) which produces a distribution over the class labels.

Many versions of DeepFace are proposed later by various researchers employing better training strategies and minor architecture changes [132, 133, 134]. Very recently, Google developed another algorithm named *FaceNet* [116] that use huge number of images for training. FaceNet used 200 million images belonging to 1 million identities for training. FaceNet achieved a record breaking accuracy of 99.63% on LFW dataset!! Another notable difference is that, unlike DeepFace and subsequent

[3] In verification, the task is to predict whether given pair of images are similar or not that is quite different from recognition

deep approaches which use classification layer, FaceNet learns an Euclidean embedding minimizing the distance between faces of same identities and increasing the distance between faces of different identities. FaceNet achieved impressive performance on face recognition and clustering tasks as well. It is clear that, training sets with millions of images really boost the performance of CNNs. However, it is a daunting task to collect such massive datasets. Omkar *et al.* [104] proposed a simple approach on how to collect and prepare such large training datasets suitable for training convnets from the Internet.

We provide a comprehensive survey of latest deep learning based approaches in Table 2.3. While the initial trend was towards designing custom architecture for face recognition, later works show superior results with general object recognition architecture such as VGG [127] or RESNET [52].

Summary: Face recognition started with semi-automatic approaches in 60's and 70's where the facial features are manually detected and then matched automatically. The data driven techniques wherein training data is used to learn an alternate representation for matching became popular in early 90s. Sparse representation classifiers gained attention few years ago, and found to be robust to noise and occlusion. These classifiers almost achieved near 100% accuracy on classical datasets with limited lighting, illumination, pose and occlusion variations, and underperform on real-world scenarios. Face recognition is currently dominated by convolutional neural networks where different approaches differ mostly in terms of model architecture, loss function and training data.

2.5 Person Recognition

While face is the most informative and reliable region, it is beneficial to take advantage of domain information, body regions such as hair, upper body or person attributes such as age, gender to improve overall recognition. This is specifically beneficial in the practical applications such as identification of players in sport videos, or in surveillance that contain additional information that help recognition.

2.5.1 With Multiple Body Regions

Multiple body regions are considered for recognition in [165]. Their approach "Pose Invariant Person Recognition" or `PIPER` consists of 109 classifiers trained on different body representations. The representations are based on CNN features trained on specific body region. They considered 107 body regions based on poselet algorithm, face and full body. The face representation is based on Deep Face [135] architecture trained on millions of images and rest are based on AlexNet [66] architecture.

Table 2.3 A brief survey of the state-of-the-art CNN based face verification approaches benchmarked on LFW dataset. The current research in face recognition is primarily focused on CNN based approaches in a supervised setting of LFW and YouTube Faces.

Method	Description	Loss	data	models	Verification metric	Layers	Accuracy
DeepFace [135]	Pre-processing involves frontalization with 2D and 3D alignment. The network has 120 million parameters and consists of two convolutional and three locally connected layers followed by a fully connected layer of 4096D.	softmax	$4M$	4	weighted χ^2	8	97.3 ± 0.25
Canonical view CNN [180]	A deep network is trained to regress a face in any arbitrary view to canonical (frontal) view. Patches are then extracted around five landmarks and passed to five small nets, followed by a joint fully connected layer.	siamese	$203K$	60	Joint Bayes	7	96.4 ± 0.25
DeepID [134]	A custom network is trained to recognize 10,000 identities. For each face, 60 features are extracted from 10 cropped regions at 3 scales, and 2 image space (color and gray)	softmax	$203K$	60	Joint Bayes	7	97.4 ± 0.26
DeepID2 [132]	An improved version of [134] with combined identification and verification loss.	softmax + siamese	$203K$	25	Joint Bayes	7	99.1 ± 0.13
DeepID3 [133]	An improved version of [132] with network architecture style of VGG Net and Google Net	softmax + siamese	$290K$	25	Joint Bayes	10	99.5 ± 0.10
Face++ [174]	Trains a standard ten-layer network on a data set of 5 million labeled faces of around 20,000 identities. With the massive training data set size, they argue that the advantages of using more sophisticated architectures and methods (such as joint verification identification loss, joint bayesian modeling) become less significant.	softmax	$5M$	1	L2	10	99.5 ± 0.36
FaceNet [116]	Learns 128D face embedding using triplet loss. The network is trained on a massive proprietary dataset with 260M images. An online hard negative mining strategy is used to train the network.	triplet	$260M$	1	L2	22	99.6 ± 0.09
VGG Face [104]	Reproduces [116] with public dataset of 2M images. A deep network is first trained for the identification of 2K identities. The features are then normalized and projected to lower-dimensional space using a triplet loss.	softmax	$2.6M$	1	L2	16	98.9 ± 0.3
dlib library [2]	ResNet network trained on 3M images derived from VGG face and Scrubs dataset.	softmax	$3M$	1	L2	152	99.3 ± 0.00

In [62], an algorithm named `naeil` is proposed that learns representations separately for multiple body regions such as face, head, upper body, body and scene. Apart from the body regions, representations based on surrogate attribute prediction tasks such as gender, age, hair color, glasses estimation. Both these authors show that recognition performance improves significantly when multiple body regions and cues are considered for challenging scenarios where face region is not reliable.

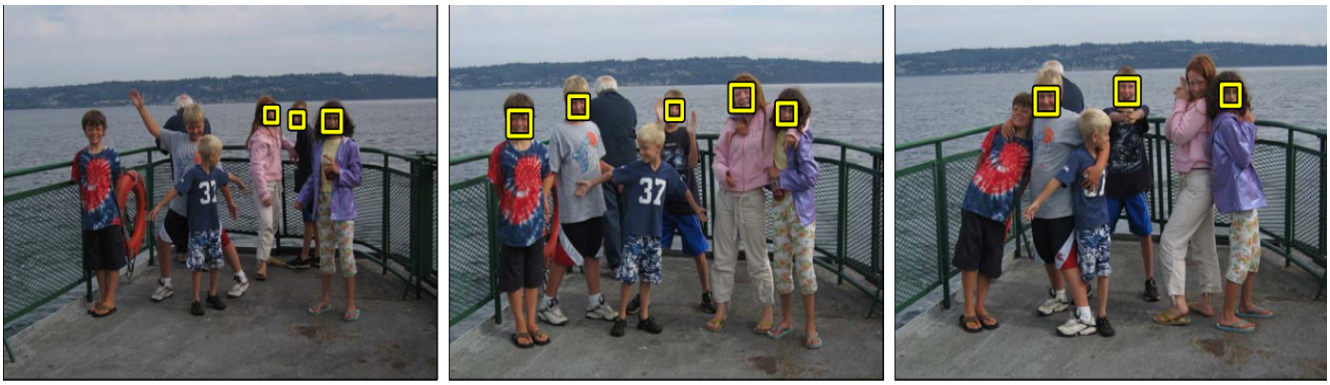
2.5.2 With Context

An interesting line of work in person recognition is to exploit application related contexts. We show few such interesting examples in Figure 2.16. In [39], the authors use timestamp, camera-pose, and people co-occurrence to find all the instances of a specific person from a community-contributed set of photos of a crowded public event. Sivic *et al.* [130] improve the recall by modeling the appearance of cloth, hair, skin of people in repeated shots of the same scene. In [7, 157], meta-data and clothing information are exploited to identify people in photo collections. Similarly, Li *et al.* [74] focused on person recognition in *photo-albums* exploiting various contexts about people and their relations in these albums. In the case of videos, multiple cues such as sub-title [29], appearance models [38, 128], clothing, audio, face [136] are exploited. Similarly, a combination of jersey, face identification and contextual constraints are used to identify players in broadcast videos [13, 90].

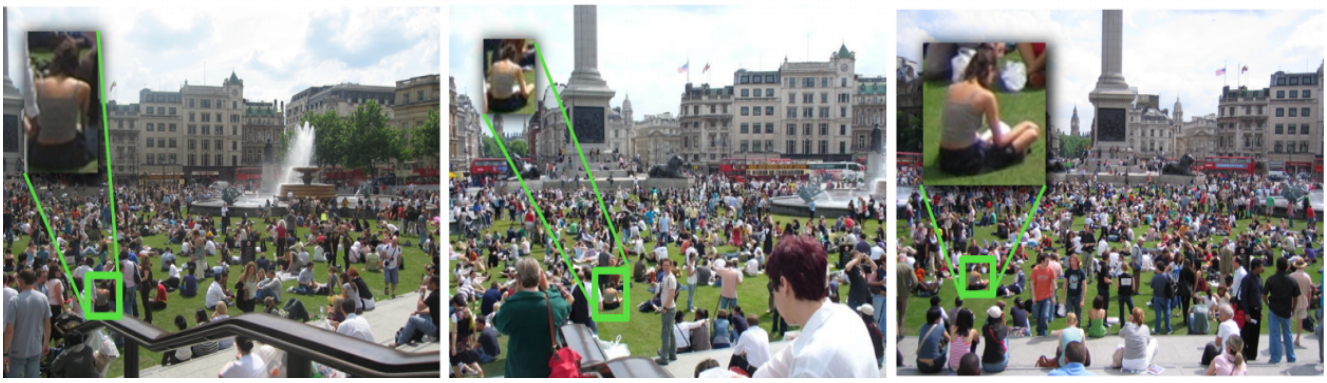
2.6 Semi-supervised Learning

Most of the algorithms described so far employ supervised learning to perform detection or recognition tasks. Supervised algorithms learn a concept from *labeled* examples during so-called "training" phase and apply that learning to make decisions on unseen examples during "testing" phase. The performance of such algorithms usually depends on the amount of labeled training data - bigger the training data, better is the performance. However, manual labeling for large amounts of data is tedious and time consuming task. It is expensive and may require expert knowledge in some scenarios such as marking abnormalities in medical image analysis.

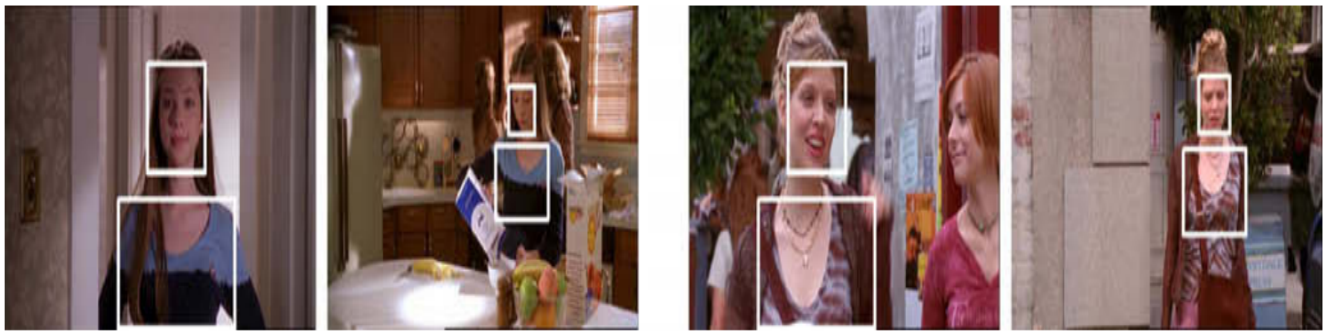
On the other hand, *unlabeled* data is abundant and relatively easy to obtain. For example, a simple text query in the Internet can produce large amount of image data. Sometimes, application of interest itself may contain or generate large volumes of data. For example, consider automatic face tagging of photos in albums or social networking sites where large amount of data is generated as users upload



(a)



(b)



(c)

Figure 2.16 Context helps in recognition. (a) shows people identification in repeated shots of the same scene [130]. The instances that are recognized correctly in some shots are used to recognize instances in other shots. (b) shows matching of people in images of the same scene [39]] using time-stamp, co-occurrence and other meta-information. (c) shows matching characters across shots uses clothing [29]] in addition to facial information.

photos. Similarly in surveillance videos or movies that contain large amounts of unlabeled data. If we develop techniques that exploit such huge amounts unlabeled data with minimal human intervention along with labeled data, then better performance can be achieved than what is possible with labeled alone. These considerations have led to an increased interest in semi-supervised learning methods that learn from both labeled and unlabeled data.

In this section, we provide a brief overview of semi-supervised approaches that use both labeled and unlabeled examples for learning. It is a broad topic and a detailed survey of these approaches can be found in [18, 175, 177].

2.6.1 Self-Training

Self-training, also known as self-learning [117, 35] is the simplest and oldest method that uses unlabeled data for classification. This is basically a wrapper algorithm that repeatedly uses a supervised learning method. It starts by training the supervised algorithm using the labeled data only. In the subsequent iterations, supervised algorithm is used to predict the labels of unlabeled examples. The supervised method is retrained using both labeled points and unlabeled points with their predicted labels. The choice of supervised algorithm is completely open. One problem associated with self-training is that if the selected supervised algorithm has a very low performance, then it may predict majority of unlabeled samples wrongly. This may deteriorate the performance rather than improving the performance when such noisy examples are used to retrain the supervised algorithm. Thus self-training should be used when the performance of the supervised algorithm is reasonably high.

2.6.2 Mixture Models

Generative learning is another popular approach that models the underlying the distribution of the given data during training and performs classification based on a decision function formed using the learned parameters. In this approach, the data is assumed to be drawn *iid* from some known distribution and the objective is to find the parameters of the unknown distribution function. If the data points are assumed to be generated from a mixture finite number of distributions, then it is referred as mixture models. While in the supervised setting where the labels of the samples are known, it is straight forward; parameters of the model can be directly estimated. However, for semi-supervised setting, since the labels are not known for unlabeled examples, an iterative approach such as EM algorithm is used for inference. In each step, EM algorithm alternates between expectation (E) and maximization (M) steps.

It estimates the hidden variables associated with each unlabeled example in the E-step and uses it to update the model parameters in the M-step. Semi-supervised generative mixture models demonstrated great success in early 2000 for text classification [100].

2.6.3 Label Propagation

Label Propagation [173, 176, 144] is another powerful graph-based semi-supervised technique. The idea is to construct a graph using both labeled and unlabeled examples and then propagate the labels from labeled to unlabeled examples using cluster assumptions. The cluster assumption ensures that the points that are nearby have similar labels. The approach is transductive[4] and involves two stages: graph construction and inference. In the graph construction step, a graph is constructed where each node corresponds to an example (either labeled and unlabeled) and an edge between them represents their similarity. The edges between similar examples should have large weights compared to dissimilar examples. The common way to compute the edge weights is using Euclidean distance based function or a Gaussian function in a neighborhood obtained using kNN or ϵ-neighborhood [173, 176]. However, for high dimensional data with large noise, one can use advanced techniques such as linear neighborhoods [173], sparse coding [68], etc for graph construction. During inference, the labels are propagated from labeled to unlabeled examples.

The main idea of label propagation is to let data samples in each iteration to absorb a fraction of label information from its neighborhood and retain some label information of its initial state in each iteration. The weights between the samples control the amount of information flow between the samples. A sample absorbs more label information from another similar sample while receive less or no information from dissimilar sample. If $F(t+1)$ define the label matrix for the entire dataset at t-th iteration and W the row-normalized similarity graph, then label matrix at $F(t+1)$-th iteration is given by

$$F(t+1) = \alpha W F(t) + (1-\alpha)Y,$$

where Y is the initial label matrix and α is a parameter that decides the amount of information retained from it initial labels. The above operation is iteratively done till convergence. It is shown in [173, 176] that the above iteration converges to $F^* = (I - \alpha W)^{-1}Y$.

[4]Transductive algorithms cannot be used to predict the labels of future or unseen examples. Both labeled and unlabeled examples are used together to predict the labels of unlabeled examples.

Chapter 3

Detecting People in Images

In this chapter, we focus on person detection which is the first step in the recognition pipeline. It involves finding the locations of one or more person body parts appearing in images. There are many possible ways to detect people pertaining to face [141, 32, 76], head [143], upper body [33], and complete body (pedestrian) [11] detections. In our work, we focus on "face detection" due to its popularity, availability of large scale training datasets and high commercial interest in the recent years in biometrics, device authentication (e.g., faceID), photo organization apps (e.g., Google Photos), surveillance, entertainment and gaming systems (e.g., camera filters), personalized marketing campaigns and so on. For systems that require the location of different body parts (e.g., multi-body person recognition), it is also possible to roughly infer their locations from face location as shown in our later chapters.

3.1 Introduction

Given an image or a video, our objective is to detect faces appearing in it. In our work, we focus on **exemplar detection framework** [76, 123]. We believe that the framework has great potential due to its (i) ability to completely avoid exhaustive sliding-window search, (ii) scalability, (iii) parallelizability and (iv) flexibility for retraining in semi-supervised setting. However, the performance of original exemplar-based approaches [76, 123] is slightly lower when compared to state-of-the-art models on challenging real-world scenarios. In this work, we aim to identify short-comings of the exemplar approach and propose solutions such that the performance gap between state-of-the-art and exemplar based approaches is reduced.

As described in the previous chapter, the existing models for face detection fall into four broad categories- Adaboost, deformable part based, deep neural network and exemplar based models. The

Adaboost method can produce real-time detections using a combination of Haar features, integral image technique and cascade of simple classifiers. Deformable part model (DPM) based techniques [179, 151, 96] model the appearance of individual object parts along with their spatial configuration to perform detection. The models obtained by training deep convolutional neural network architecture on large face datasets are performing extremely well in the recent years. All these aforementioned approaches distill compact models of faces from large training database in order to capture most common variations in pose, expression, lighting, *etc*.

On the other hand, exemplar-based detectors [76, 123] do not explicitly model the face variations and is getting quite a lot of attention lately. The approach combines bag-of-words (BoW) retrieval technique and Hough voting [36, 73] to perform efficient detection. In this paradigm, a large database of exemplars that cover significant face variations are collected. Local features (such as SIFT) are extracted, quantized and indexed using traditional BoW technique. For detection, each exemplar casts a vote on the given target image at multiple scales after which the votes are aggregated. Since each exemplar is *specific* to particular variation, it is possible to detect faces in challenging conditions using a sufficiently large database with diverse exemplars.

Existing exemplar approaches [76, 123] treat each exemplar as a collection of independent visual words that capture facial features from different regions. It is however apparent that many visual features co-occur in faces. For e.g., stable visual features that describe eyes and nose occur together with greater probability. Thus, the current exemplar schemes fail to capture such semantic relations among visual features, unlike model-based approaches, which are designed to capture higher order spatial relations. We propose to incorporate such higher order information using **"visual phrases"** in the exemplar framework. Visual phrase is a group of highly correlated and stable visual words that co-occur in faces frequently. We discover such visual phrases from an exemplar database. As it is computationally expensive to model all possible dependencies for a large vocabulary, we employ a popular association rule mining technique [4] and obtain large candidate visual phrases that occur frequently. We then retain only those phrases that are suited for detection through a discriminative training process.

We also introduce a domain-specific similarity function that considers the **"spatial context"** of visual features to improve their discriminative ability. This is in contrast to non-discriminative inverse document frequency (IDF) based function used in current schemes that ignore such spatial information. Our approach is based on the observation that a stable visual word or a phrase appears at *consistent locations* and in *consistent exemplars*. We leverage the availability of a large database to estimate the

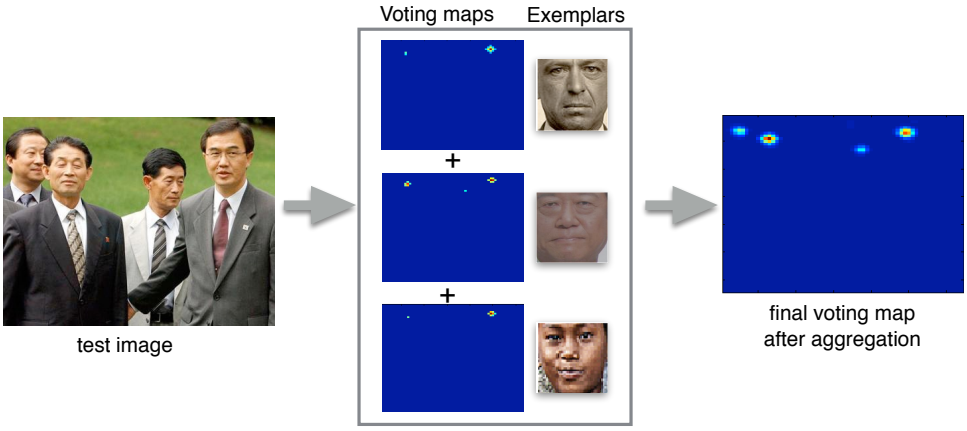


Figure 3.1 Ensemble of Exemplars for Face detection: A large database of diverse exemplars is collected, and indexed using a BoW representation. During testing, each exemplar casts a vote on the test image at multiple scales. The votes from different exemplars are then aggregated to detect the faces.

spatial distribution of words and phrases and weigh their individual occurrences in each exemplar based on this distribution. This ensures that visual words and phrases in exemplars cast a strong vote only if they occur at their globally consistent locations. This suppresses the contribution of noisy features introduced due to imperfect feature extraction and quantization processes.

Some of the works in the area of content-based retrieval use similar insights. In [22, 58, 124], visual word dependencies in a database with multiple objects and scenes are discovered. While such dependencies are suppressed for retrieval tasks [22, 58, 124], we exploit them as positive cues for detection. In [146], the contextual weighting of the features is proposed but for sparse local features. The work of Yuan *et al.* [161] is closely related to ours. They demonstrate an approach to discover meaningful *visual phrase lexicons* with spatially consistent visual words given a large database. Visual phrases are also applied in image retrieval [168, 167], object recognition [167] and detection [114] tasks.

3.2 Exemplar Framework for Face Detection

In this section, we give a brief overview of exemplar framework which forms the basis of our work. In the exemplar framework [123, 76], local features such as dense-SIFT are extracted from a large exemplar database and a k-means based vocabulary is constructed followed by feature quantization.

Term frequencies (TF) and inverse document frequencies (IDF) are calculated and inverted files are created. The training procedure is very much similar to BOW retrieval scheme [129].

During testing, all the exemplars collectively participate in the Hough-based voting [36, 73] process that uses the spatial locations of features to locate the faces in a given image. Each exemplar generates a voting map at multiple scales, where each location in the map indicates the similarity score between the exemplar and the image sub-region at that location as shown in Figure 3.1. The similarity measure between an exemplar e_i and the rectangular region centered at location p of the test image x is given as [122, 123]:

$$S(p, e_i) = \sum_{k} \sum_{\substack{f \in R_x(p), g \in e_i \\ w(f)=w(g)=k \\ ||\mathbb{T}(L(f))-L(g)||<\epsilon}} \frac{F(w(g), L(g))}{tf_{e_i}(k) \cdot tf_x(k)}, \tag{3.1}$$

where x is the test image, $R_x(p)$ is the sub-image region of x centered at p. f and g are the local features and $L(f)$ and $L(g)$ are their corresponding locations from x and e_i, respectively. $w(f)$ and $w(g)$ are the quantized visual words of features f and g respectively. $w(f) = w(g) = k$ indicates that only the matched visual words are considered for voting. The spatial constraint $||\mathbb{T}(L(f)) - L(g)|| < \epsilon$ further ensures that matched features should be closer under some unknown transformation $\mathbb{T}$. $F(w(g), L(g))$ is the weightage given to each matched feature pair quantized to visual word k. To handle burstiness, weights are divided by $tf_{e_i}(k)$ and $tf_x(k)$, which denote the TF of the visual word k in the exemplar and test image, respectively [122].

Voting maps can be efficiently computed using generalized Hough-voting [36, 73] to detect faces of various sizes. Suppose that we are interested in detecting faces of size $N_x \times N_x$ in the test image[1]. The location p where the vote is cast is calculated as follows.

$$p = L(f) + \frac{N_x}{N_{e_i}}(C_{e_i} - L(g)), \tag{3.2}$$

where C_{e_i} and N_{e_i} are the center and size of the exemplar e_i, respectively. This scheme completely avoids the exhaustive sliding-window mechanism followed in model-based approaches for detection. Finally, gating is applied on each of the voting maps by subtracting the scores with a pre-trained threshold. The resulting maps from different exemplars are then aggregated to obtain the final voting map as. The voting maps are then subtracted with an exemplar specific threshold and aggregated to obtain the

[1] with an aspect ratio of 1:1 for exemplars and target faces.

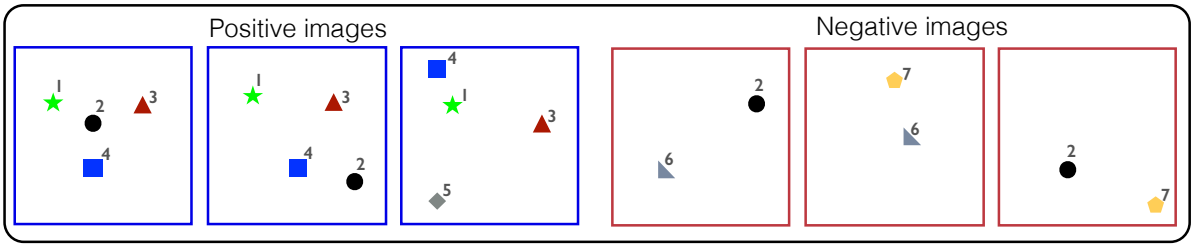


Figure 3.2 Motivating reasons: Consider three positive and negative examples. Current schemes that ignore spatial location assign an high IDF weight for words 1, 3 and 4. However, word 4 in third positive example is occurring at inconsistent location possibly due to noise that should be given slightly less weightage. Also, word 2 appears in both positive and negative examples and should be given less weightage. Word 4 in positive example 3 do not occur at its globally consistent location, hence should contribute less.

final voting map [123]:

$$S(x) = \sum_{i:s_i(x)>\rho_i} (s_i(x) - \rho_i), \tag{3.3}$$

where $s_i(x)$ is the similarity score between x and e_i, and ρ_i is the discriminatively trained threshold for exemplar e_i obtained during training.

3.3 Contextual Weighting of Features

Current exemplar detectors compute the similarity scores between the exemplar and a target image sub-region as [123],

$$F(w(g) = k, L(g)) = idf^2(k), \tag{3.4}$$

where $idf(k)$ is the IDF of the visual word k. The voting scheme with above similarity score has two issues. First, the use of IDF computed from only the positive exemplars makes it less discriminative for detection tasks. Second, the approach assumes that exemplar words are noise-free and considers all the visual words equally important when computing the similarity score. However, a noisy feature that is wrongly assigned to a visual word with high IDF may significantly affect the voting process.

Figure 3.2 illustrates these issues with a simple example with 3 positive and 3 negative exemplars. Current exemplar approaches consider only positive exemplars and will give a high IDF to vocabulary elements $1, 3$ and 4. This will also assign a high IDF to the vocabulary element 2, even though it occurs with similar probability in both positive and negative images. Another issue is that, a highly discrim-

Figure 3.3 Spatial Context of visual words: (a) and (b) shows the location of two visual words in different images. Notice how the visual word in (a) is *highly localized* with consistent locations while the word in (b) appears at random locations (left). The global distributions of each visual word over the entire database (middle) is used to weight their occurrences in individual exemplars. Its overlay on the mean exemplar face (right), shows strong localization for stable words. Unstable words occurs at diverse locations and are down weighted.

inative word occurring at an incorrect location in an exemplar may cast a wrong vote. In Figure 3.2, the visual word 4 is discriminative as it occurs in consistent locations in positive exemplars 1 and 2. However, a feature in exemplar 3 is wrongly assigned to visual word 4 due to noise, and if we ignore its location, may contribute incorrectly during voting.

Motivated by these observations, we address the following questions: How can we down-weight less discriminative vocabulary elements? and How can we discover noisy features in exemplars and down-weight their contribution during voting? Our modification is based on the argument that a visual word that is stable and discriminative tends to occur consistently in *similar locations* and in *similar exemplars*. Similarly, a visual word that is noisy or less discriminative with very high probability occurs at random locations. This is illustrated in Figure 3.3, where a stable visual word that describes the appearance of nose in a particular view (here frontal) appears consistently at the same location in other similar exemplars, or in other words, it is highly *localized*. We estimate the distribution of each visual word from the entire database and use it to weight their occurrence in exemplars. Based on this, visual words appearing at their globally consistent location get more weightage while those appearing at random locations get less weightage.

Let, $w(L(g), e_i)$ denote the visual word corresponding to feature g at location $L(g) = (L^{\mathrm{x}}(g), L^{\mathrm{y}}(g))$ in the exemplar e_i. We estimate the distribution of each vocabulary element k from

the entire exemplar database as,

$$P_e(k|(\mathtt{x},\mathtt{y})) = \frac{1}{N_e}\sum_{e_i}\mathcal{I}(w(L(g),e_i) == k), \tag{3.5}$$

where N_e denote total number of exemplars, $(\mathtt{x},\mathtt{y})$ denote the location and $\mathcal{I}(\cdot)$ is an indicator function whose value is 1 if the condition is satisfied, otherwise 0. In the practical cases, however, there will be some misalignments between the exemplars.

To handle the misalignments, we convolve the distribution with a 8×8 Gaussian filter H ($exp(-d/\sigma^2)$ and $\sigma^2 = 2.5$) to obtain the spatial weightage for each vocabulary element as,

$$W((\mathtt{x},\mathtt{y}),k) = P_e(k|(\mathtt{x},\mathtt{y})) * H \tag{3.6}$$

We show such weightage obtained for a stable and non-stable word in Figure 3.3. It also suggests to suppress the unstable words that would otherwise affect the voting process. Similarly, we estimate distribution of each vocabulary element on a large corpus of negative images n_i as

$$P_n(k|(\mathtt{x},\mathtt{y})) = \frac{1}{N_n}\sum_{n_i}\mathcal{I}(w(L(g),n_i) == k), \tag{3.7}$$

where N_n denote the total number of negative images. We then compute the global *discriminative score* $\mathcal{D}$ for each visual word k as,

$$\mathcal{D}(k) = \max_{(\mathtt{x},\mathtt{y})}\frac{P_e(k|(\mathtt{x},\mathtt{y}))}{P_n(k|(\mathtt{x},\mathtt{y}))} \tag{3.8}$$

The above score $\mathcal{D}(\cdot)$ is discriminative and also considers the spatial location of features, hence is suited for detection. Finally, our scoring function for every matched feature pair between exemplar and target sub-region is given as,

$$F(w(g) = k, L(g)) = W(L(g),k)\cdot\mathcal{D}^2(k), \tag{3.9}$$

where $W(\cdot)$ denote the context-aware weightage given to exemplar feature and $\mathcal{D}(\cdot)$ is the discriminative score.

3.4 Visual Phrases for Detection

Due to the independent assumption in previous exemplar approaches, each visual word independently votes for the target image. However, for faces, it is intuitively obvious that many visual words are *highly correlated* and *co-occur* together. The current schemes fail to capture such semantic relations

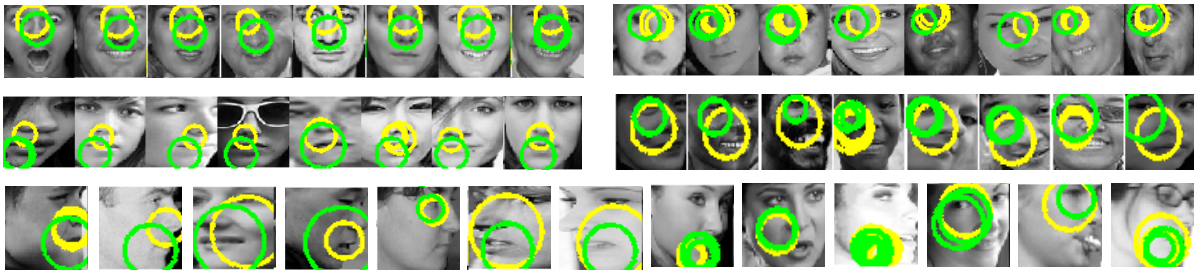


Figure 3.4 Visual Phrases for Faces: The top two rows (left and right) shows 4 different visual phrases that capture relation among two visual words. Notice how the stable visual phrases capture semantic relation among different visual features. Visual phrases are highly localized and appear at similar locations in similar exemplars. Bottom row shows few other visual phrases discovered from the database.

among visual features, unlike in model-based approaches that capture much complex relations. Though the terms in the denominator of Eqn 3.1. handles burstiness, it does not consider the relation among the visual words.

We propose to incorporate higher order information using so called visual phrases in the exemplar framework. A visual phrase is a group of spatially consistent and semantically related visual words that co-occur in faces. We leverage the presence of large database to discover such visual phrases. Given a large vocabulary, it is however, computationally expensive to find all such dependencies. To this end, we resort to a popular data mining technique, *association rule mining* [4] to obtain the candidate visual phrases that occur frequently in the database. We then prune the candidate set and retain only those visual phrases that are well suited for detection.

It is worth to note that, such relations are earlier exploited in computer vision for retrieval tasks [22, 58, 124]. In these tasks, images usually contain multiple objects and scenes and a similarity function with independence assumption tends to over-weight the regions containing highly correlated words [22, 124]. Therefore such correlated words are down-weighted for better retrieval. However, we exploit such relations among visual words for detection as they provide strong cues about the existence of a face region.

We now formally discuss the proposed approach to discover visual phrases. Let $V = \{v_1, v_2, \ldots, v_n\}$ denote the vocabulary and e_i be the exemplar containing subset of vocabulary elements i.e. $e_i \subseteq V$. An association rule [4] is an implication of the form $X \implies Y$, where X and Y are the itemsets (*visual phrases*) that satisfy $X \subset V$, $Y \subset V$ and $X \cap Y = \varnothing$. The implication rule basically checks with

what proportion the itemsets X and Y occur together in an image e_i. The result is a list of all possible combination of words with a support[2] greater than user-specified threshold.

The candidate visual phrase set obtained from the above algorithm on a large database is usually huge containing many redundant phrases. It may also be possible that many of the visual words occur together by chance. Also, the mining technique does not consider the spatial location of words due to which many of the candidate visual phrases are not discriminative for detection task. Due to these reasons, we need to prune the candidate phrases obtained from the rule mining and select only those discriminative phrases that are suitable for detection. We achieve this using the concept of spatial consistency introduced earlier for visual words. We consider the visual phrase as stable and discriminative if all the words associated with it appear in consistent locations in the exemplars, and occur rarely in negative images.

Let, $\Omega = \{\eta_i \mid \forall i, \eta_i \subset V\}$ be the list of candidate visual phrases discovered from association rule mining and $|\eta_i|$ denote the number of words associated with the visual phrase η_i. We assign a score for each candidate visual phrase η_i as,

$$\mathcal{Q}(\eta_i) = log\Big(\frac{1+\Psi^+}{1+\Psi^-}\Big), \tag{3.10}$$

where

$$\Psi^+ = \max \sum_{\substack{\forall k, k \in \eta_i \\ \forall (x,y)}} P_e(k|(x,y)) * H$$

$$\Psi^- = \max \sum_{\substack{\forall k, k \in \eta_i \\ \forall (x,y)}} P_n(k|(x,y)) * H$$

The terms Ψ^+ and Ψ^- measure the spatial consistency of the words that constitute visual phrase in positive and negative images, respectively. The score $\mathcal{Q}$ in Eqn 3.10 will be large for those visual phrases that capture the relation of stable visual words, and less for non-discriminative and noisy phrases that occur at random locations. We finally retain the visual phrases whose $\mathcal{Q}$ score exceeds a threshold i.e. $\omega = \{\eta_i \mid \mathcal{Q}(\eta_i) > \rho\}$ (see Section 3.6.1). We show few visual phrases discovered from the exemplar database in Fig 3.4. Notice how the visual phrases capture the neighbourhood (spatial and scale) relations due to multi-scale dense feature extraction (e.g., bottom row 5th and 8th image). Once the phrases are discovered, we index their occurrences in exemplars and incorporate them into the voting framework. The spatial location $L(\cdot)$ and discriminative score $\mathcal{D}(\cdot)$ of the selected visual phrases

[2]Support is the number of transactions (images) in the database that contain the itemset (phrase) or simply the frequency count of a phrase.

($\eta_i \in \omega$) are obtained using the mean location of visual words and sum of their individual discriminative scores, respectively.

$$L(\eta_i) = \frac{1}{|\eta_i|} \sum_{\substack{w(g)==k \\ \forall k, k \in \eta_i}} L(g) \tag{3.11}$$

$$\mathcal{D}(\eta_i) = \sum_{\forall k, k \in \eta_i} \mathcal{D}(k) \tag{3.12}$$

3.5 Time and Memory Complexity

When compared with baseline exemplars, the proposed approach requires additional memory for indexing the visual phrases and the contextual weights of visual words and phrases. The average number of visual phrases discovered per exemplar was around 8. In the case of a database of 15k images, this results in an additional memory of 1MB to index the visual phrases and their locations following the representation in [122]. The contextual weights are quantized and stored using 1 byte integer that requires additional 10MB of memory for 80×80 exemplar with 700 visual words and phrases on an average. One could reduce the memory footprint by removing those visual features and phrases with very low contextual weights as they make limited contribution in the voting process. As we demonstrate later, it is possible to remove upto 30% of features with a slight drop in performance. Compared to previous approaches, the only additional time required is to find the dependencies in the target image. Since the target image is indexed and TFs are computed already for voting process, dependencies can be found much faster. This usually takes less than 2 seconds in our unoptimized MATLAB code. Also, similar to [122], we can achieve a further speedup by ignoring the words with high TFs, as their contributions are limited according to Eqn 3.1. Our MATLAB implementation of the entire detection pipeline without tiling [76] usually takes $10 - 12$ secs for 1280×1280 image most of which is spent in feature extraction and quantization. The source code is available at `http://cvit.iiit.ac.in/projects/exemplar/`.

3.6 Experiments and Results

We implemented the exemplar detector [123] upon which our improvements are made. The performance of our baseline exemplar closely matches with [123] as shown in Fig 3.8.

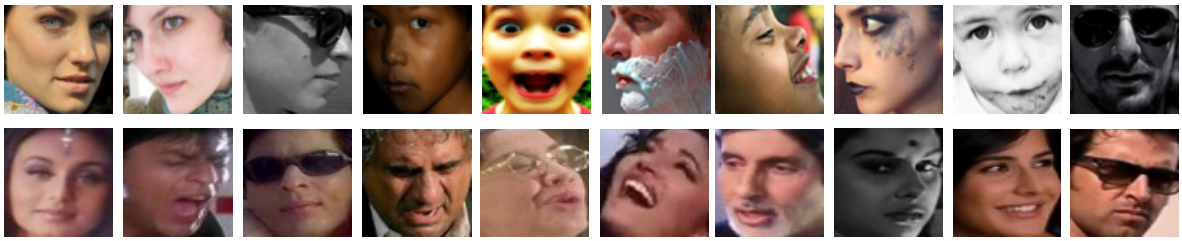


Figure 3.5 Few images from our database built from AFLW (top) and IMFDB (bottom).

3.6.1 Implementation details

Exemplars: We collected the exemplar images from AFLW [65] and IMFDB [119] databases. AFLW contains around 25k images and IMFDB contains 34512 images. We randomly sample 10k images from AFLW and 5k images from IMFDB to create our 15k exemplar database. All the exemplars are resized to a fixed size of 80×80. Few exemplars from our database are shown in Fig 3.5.

Dense Features and Vocabulary: We densely extract patches of size 24×24 with a stride of 3 pixels and 128D root-SIFT representation is computed. We extract the features and their locations at 12 scales by resizing the original image with a scaling factor of $\sqrt{2}$. We construct the 50k-vocabulary using fast approximate nearest neighbour (ANN) k-means [97]. We used the publicly available software VLFEAT [140] for both these tasks.

Visual Phrases: We use the publicly available apriori software [1] to obtain the initial candidate phrases with a minimum support of 100. This resulted in 5837 candidate visual phrases containing 4880 2-visual word phrases, 736 3-visual word phrases and 221 4-visual word phrases. We used 50k 80×80 negative patches [24] with a threshold of $\rho = 0.5$ for discriminative training that finally resulted in 1282 2-visual phrases. Few visual phrases are shown in Fig 3.4. We noticed that 3 and 4- word phrases were noisy and inconsistent ($\mathcal{Q}(\cdot) < \rho$) and hence are not considered.

Voting and Thresholds: We considered a voting map of size 64×64 similar to [123, 76] and obtained the corresponding grid size using smallest image dimension. To avoid quantization errors, maps are smoothened using a 5×5 Gaussian filter $exp(-d/\sigma^2)$ and $\sigma^2 = 2.5$. The gating threshold for each exemplar is obtained by selecting the maximum score on 1000 negatives images [24] when voted using the same exemplar [123].

Detection: For better performance, test images are upscaled to have a size of atleast 1280 [123]. For memory efficiency, we follow tile-based detection and divide the upscaled image into tiles of size

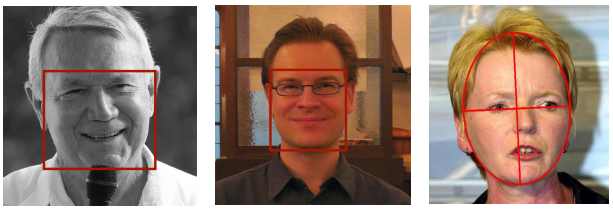


Figure 3.6 Annotation mismatch: Notice the difference annotation strategies across (a) AFLW [65] (b) AFW [179] and (c) FDDB [56].

640×640 with an overlapping stride of 140 pixels [76]. The detection operation on each tile is performed at 3 different scales $(1, 0.5, 0.3)$. At each scale, faces of 15 sizes with a base size of 80×80 and scaling factor of $2^{1/4}$ are detected. We vote only using top 3000 similar exemplars retrieved using BOW model to speed up the processing. A standard greedy non-maxima suppression (NMS) with an overlap threshold of 0.25 is applied to suppress overlapping detections.

Bounding box adjustments: Different face detection benchmarks have followed different annotation strategies (see Fig 3.6). Due to this mismatch, the algorithms that are tuned for particular type of annotations seem to perform badly for other datasets that have followed a different annotation strategy. As in previous works [76, 96, 153] we also modify the detected face regions to better match the location and scale of the ground truth bounding boxes. For example, for the FDDB dataset [56], we convert a detected bounding box of size (w, h) to vertical ellipse with parameters $(\frac{0.9w}{\sqrt{2}}, \frac{h}{\sqrt{2}})$.

3.6.2 Datasets

We show our results on popular face detection benchmarks - FDDB [56] and AFW [179] and G-album [37]. All these datasets offer challenging scenarios for face detection since the images are collected from various Internet sources. FDDB [56] contains a total of 5171 faces in 2845 images collected mainly from Yahoo news website. The dataset contains very low resolution images (smaller than 30×30) that truly tests the capability of algorithms. We use the ROC evaluation software that comes with FDDB database as recommended by the dataset creators and commonly followed by researchers. AFW [179] contains 468 faces present in 205 images. The database is characterized by cluttered background with pose, aging, and occlusion variations. G-album [37] dataset contains 589 family photos with 931 faces. The images contain natural outdoor scenes with strong pose, expression, illumination

variations offering a more realistic scenario. We compare our results on AFW and G-album datasets in terms of precision-recall (PR).

3.6.3 Results

We compare the performance of the proposed approach with the previous exemplar schemes [123, 76] in Fig 3.9. We consistently outperform previous schemes on both FDDB and AFW datasets. Following FDDB protocol, we compare our results with all the previously published results in Fig 3.10. From the discrete curve in Fig 3.10(a), it is clear that our proposed approach, not only improves over exemplar schemes but also outperforms most of the previous non-exemplar schemes [152, 179, 78, 77, 41], except [96]. The contributions of contextual weighting and visual phrases to the performance improvement is given in Fig 3.8. While context helps to suppress noisy inconsistent features, visual phrases complement it with its ability to upweight the co-occurrence of visual words in faces. Thus a combination of the two approaches indeed helps as can be observed from Fig 3.8. We also show the continuous curve in Fig 3.10(b) that measures the bounding box overlap with ground truth. Unlike [151, 96] that fits oriented bounding boxes, we fit a vertical ellipse which results in a slightly lower score.

For AFW, we used the evaluation software [96] to compensate for bounding box misalignments by learning a transformation for each algorithm that maximizes the detection box and ground-truth overlap. Fig 3.11(a) shows the comparison of our approach with several academic (TSM [179], DPM, HeadHunter and SquaresChnFtrs [96] and Structured models [152]) and commercial solutions (face.com, Face++, Google Picasa). Our approach achieves very high performance reducing the gap between DPM and exemplar based approaches. The common reasons for failure are bounding box misalignment, extreme pose and low resolution. For images with extreme poses and low resolutions, lack of informative features around discriminative regions such as eye and nose causes exemplars not to match unlike holistic matching methods (see Fig. 3.13). While the methods proposed in [96] benefits from decades of research along with refined implementations, exemplar approach is still in its earlier stages and we believe that the performance will be further improved in the future. Finally, we show the performance of our approach on G-album dataset. For this dataset, we compare with baseline exemplar [123] and DPM [96] using their trained model. Our approach not only improves upon exemplar method but matches the performance of DPM on this dataset as shown in Fig 3.11(b). As discussed earlier, it is possible to save memory by removing less consistent features using contextual weights. As shown in

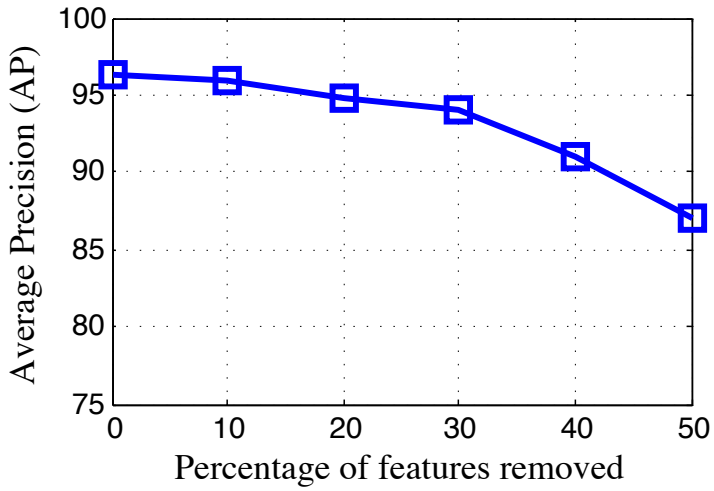


Figure 3.7 Merits of contextual weighting. It is possible to remove upto 30% of features without a significant performance change.

Fig 3.7 for the AFW dataset, it is possible to remove up to 30% of features without a significant drop in performance.

Since the publication of our work, many works have been published, majority of them using deep features and achieved better results. We provide a comparison of our approach with latest research in Table 3.1. These results suggest that it is possible to improve exemplar framework by replacing shallow SIFT features with local convolutional features [8, 99]. Our initial attempt with deep features do not show encouraging result as they are generic in nature due to the training procedure. As the models trained for classification or detection tasks remove intra-class variations, the resulting features are more generic, and are unable to distinguish between different facial variations in exemplar framework. Another possible direction is to use region proposal framework in which face proposals are first generated and validated using exemplar voting. We hope to investigate these directions in the future.

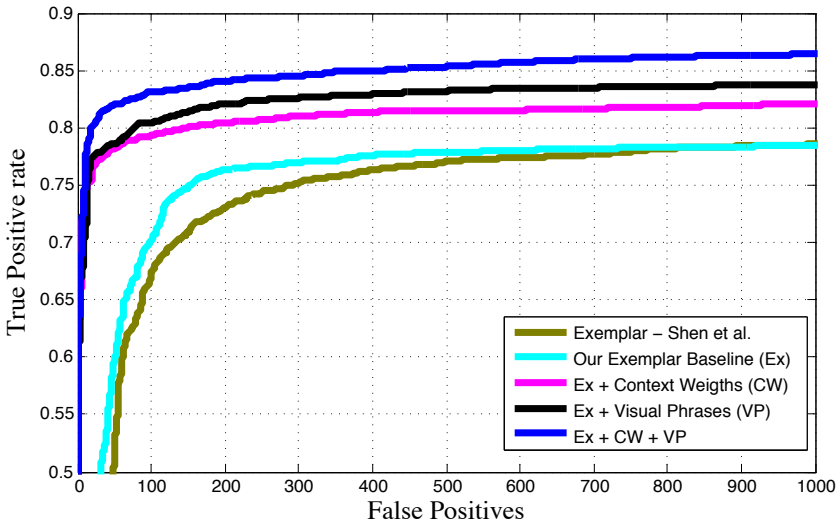
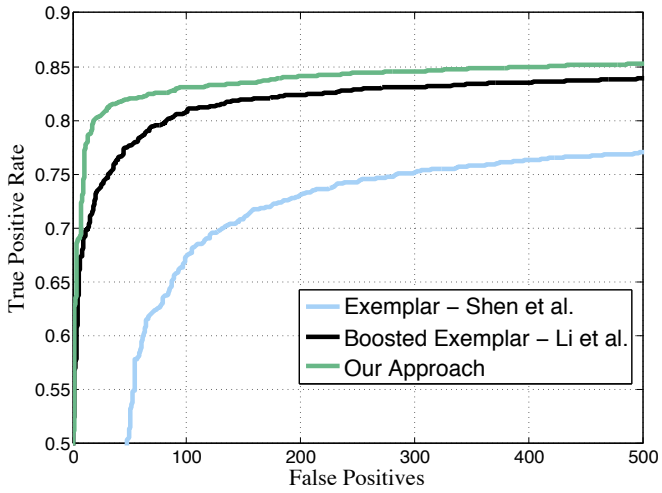


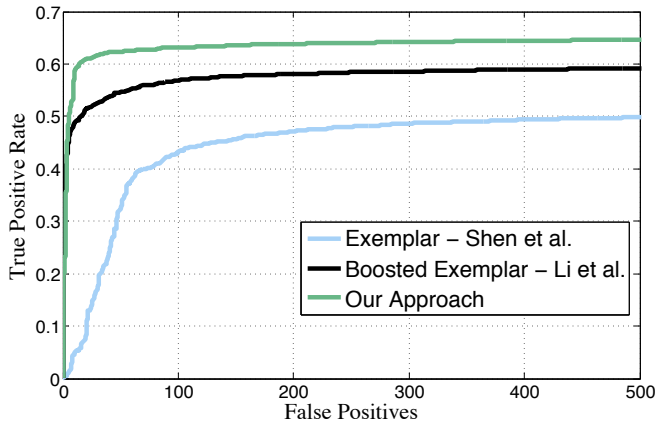
Figure 3.8 Role of contextual weights (CW) and visual phrases (VP) in improving the performance of exemplar detectors on FDDB dataset.

Table 3.1 Detection rates of different approaches on FDDB dataset measured in terms of true positive rate (%) @ 2000 false positives.

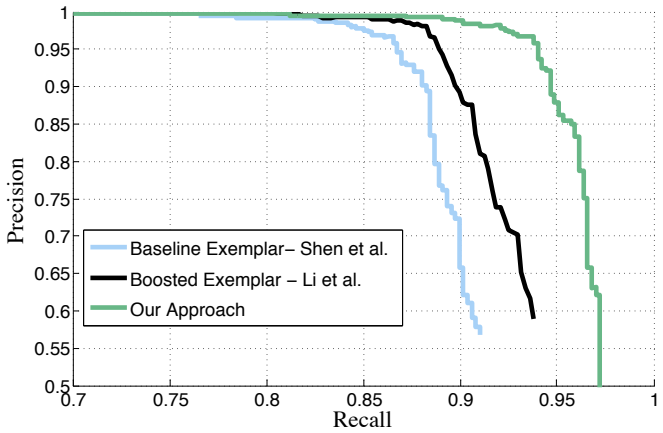
Method	TP @ 2000 FP (%)
Viola Jones [84]	59.7
Exemplar w/o contexts & visual phrases [76]	77.7
SURF cascade [78]	83.7
Integral Features [153]	85.7
Cascade CNN [77]	85.7
DPM [96]	86.4
Ours- Exemplar with context & visual phrases	86.4
Head Hunter [96]	87.1
Hyper Face [109]	90.8
Faster RCNN [60]	96.1
Tiny Faces [54]	98.3
Face R-FCN [147]	99.4



(a) FDDB Discrete ROC curve

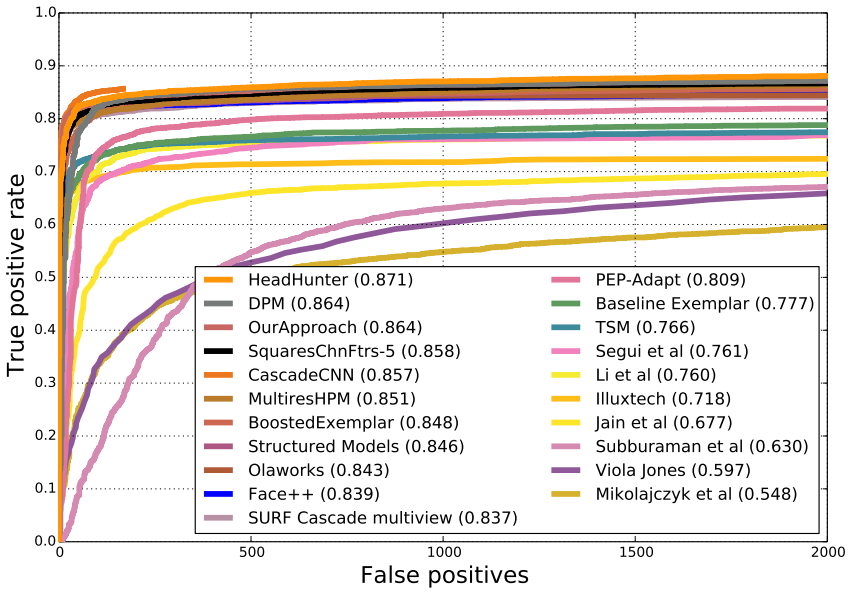


(b) FDDB Continuous ROC curve

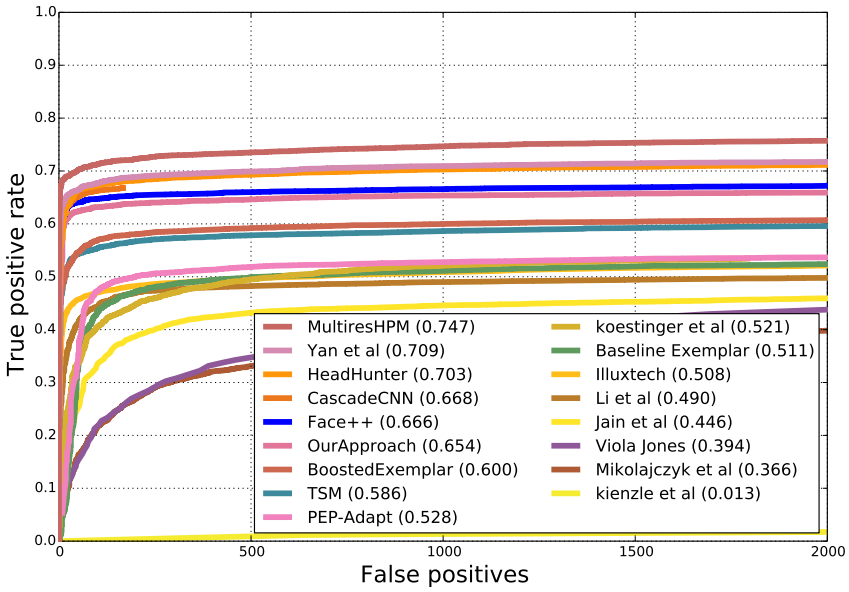


(c) AFW PR curve

Figure 3.9 Comparison with previous exemplar schemes. We outperform the baseline Exemplar [123] and Boosted Exemplar [76] on both FDDB ((a) and (b)) and AFW (c) datasets by a large margin.

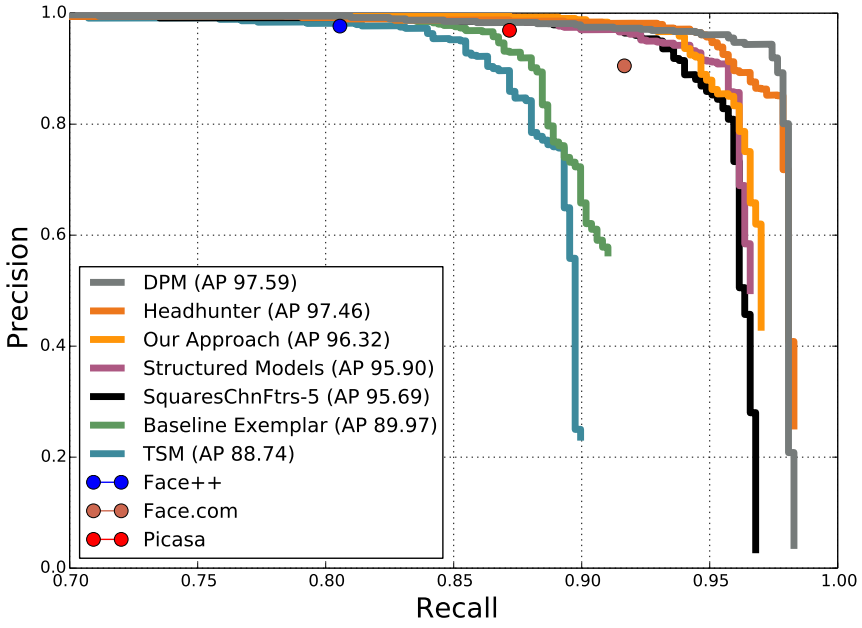


(a) Discrete ROC curve

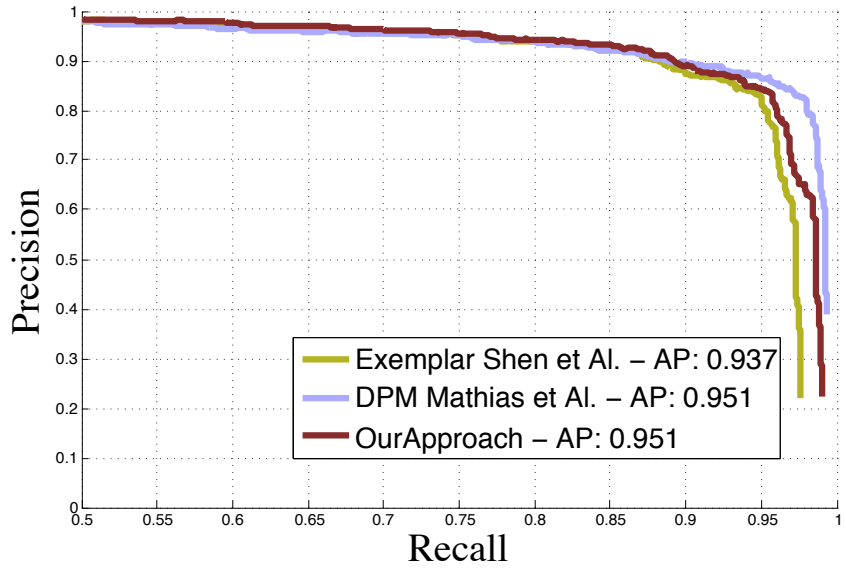


(b) Continuous ROC curve

Figure 3.10 Comparison with other approaches on FDDB dataset. We achieve an average precision of 86.4% with a negligible difference compared to HeadHunter [96]. Our performance improves over the baseline exemplar approach [123] by almost 8%.



(a) AFW



(b) G-album

Figure 3.11 Performance comparisons on AFW and G-album datasets. While our approach achieves superior performance on AFW compared to many academic and commercial approaches closely matching HeadHunter [96], the performance matches DPM [96] on G-album dataset.

Figure 3.12 Qualitative results of our detector over FDDB (top), AFW (middle) and G-album (bottom).

Figure 3.13 Failure cases. Due to bounding box misalignments (a) detected faces in green are considered false positives. (b) and (c) lack informative features due to extreme pose and low resolution.

3.7 Summary

In this chapter, we focused on face detection that forms the pre-requisite for a recognition system. We aimed at exemplar framework and introduced visual phrases to capture the semantic relations among the visual words and propose a method to incorporate them into exemplar framework. We also introduced the notion of spatial context into the framework that weights different visual words based on their prior distribution estimated from the exemplar database. Our domain-specific similarity score considers both spatial consistency and discriminative ability of visual words and phrases, and hence is suited for detection tasks. We showed from our experiments that incorporating visual phrases and contextual weights can significantly improve the performance of exemplar detectors on various face detection benchmarks.

Chapter 4

Recognition with Limited Training Samples

4.1 Introduction

In this chapter, we turn our towards "person recognition" and address the limitation of training examples in practical systems. The data unavailability arise mainly due to two reasons. First, it is not feasible or possible to capture multiple images of the subjects in many practical recognition systems. For instance law enforcement, biometric or surveillance systems usually have a few user photos collected from various identification sources (such as *Aadhar*, passport, social security number, driving license, *etc*). Second, even in applications that have access to large number of images (such as photo-album organization or celebrity identification in videos), it is very expensive and time consuming to manually vet and label large number of images.

On the other hand, unlabeled data is abundant, cheap and relatively easy to obtain. For example, a simple text query in the Internet can produce large amount of image data. Sometimes, systems itself has access to large volumes of unlabeled data. For example, consider automatic face tagging of photos in albums or social networking sites where large amount of data is generated as users upload photos. Similarly surveillance videos or authentication systems might generate large unlabeled data of the subjects to be recognized over a period of time. In all such situations, it is effective to adopt techniques that exploit such huge amounts of unlabeled data with minimal human intervention, and produce better results than what is possible using labeled data alone. These practical considerations have led to an increased interest in semi-supervised learning methods [175] that learn from a small amount of labeled data and a large amount of unlabeled data.

In this chapter, we propose a solution for practical person recognition that overcomes the limitation of labeled samples by leveraging unlabeled samples using a semi-supervised framework. Note that, we

recognize people from their "faces", and assume that the faces are already detected and cropped, to keep our attention primarily on the core issue of limited training samples. We highlight that the approach is generic and can be extended to other forms of person recognition.

Our proposed solution combines two kinds of semi-supervised learning, namely *self-training* and *label propagation* algorithms. Self-training is an iterative algorithm that hypothesizes the labels for the most confidently classified unlabeled samples. These predicted labels are then considered as additional labeled training data. In each iteration of the self-training algorithm, we use label propagation framework [176] to predict the labels of the unlabeled samples. It is a graph based framework that produces class scores based on the similarities of different samples. The choice of similarity measure is critical for label propagation. As traditional similarity measures based on Gaussian or Euclidean distances are not robust enough for face recognition, we develop algorithm named **"nearest neighbor based sparse coding (NNSC)"** to obtain the graph weights or similarities. NNSC inspired from [145, 150], represents each sample as a linear combination of its nearest neighbors. It is much faster with performance similar to l_1-based sparse representation [150], a popular representation for faces, for obtaining graph weights. NNSC representation ensures that positive edge weights are formed within neighborhood of each point unlike sparse representation that forms connection between far away points.

Finally, we show the extension of our algorithm to predict new test instances samples using NNSC representation and reconstruction error based classifier. After obtaining the representation of the query image using NNSC, it is assigned the label of the class that produces minimum reconstruction error. We conduct extensive experiments on classic face recognition datasets such as Yale [72], AR [94] and CMU PIE [125] and demonstrate the performance improvement obtained with our proposed approach.

We briefly review the works that have previously dealt with lack of training samples. In [107], self-taught learning whose goal is to learn unsupervised feature representation for classification. A dictionary is trained using unlabeled samples alone in l_1-based dictionary learning framework. Labeled samples are then represented sparsely over the trained dictionary and linear SVM is trained to predict the class labels. Self-training based approaches are used for face recognition with few training samples [113, 172, 17] with their underlying prediction algorithms based on dimensionality reduction techniques. In [113], Eigenspace (PCA basis) is computed using the labeled samples. A template for each class is then created using the mean projected representations of the labeled samples belonging to a particular class. Then, the unlabeled samples that are closer to the projected mean templates of each class are selected and augmented to the labeled set and the procedure is repeated till all the unlabeled samples

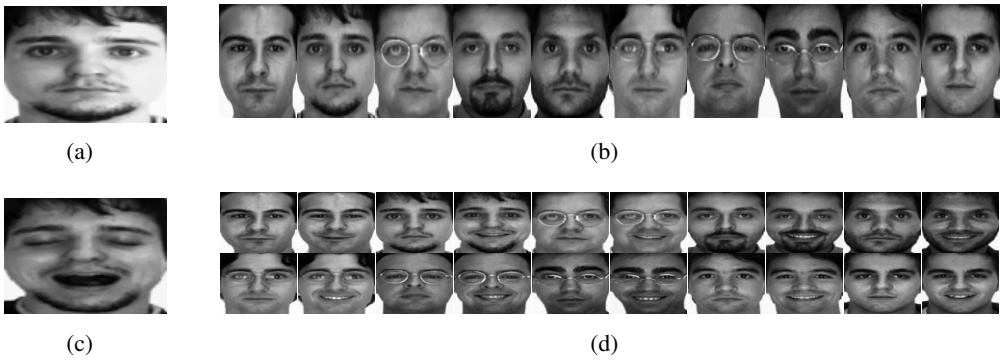


Figure 4.1 Demonstration about the effect of lack of training samples. SRC identified the (a) first image correctly but failed to identify the (c) second image with (b) small dictionary d_1 (one sample per subject). Using (d) large dictionary (two samples per subject), it identified both images correctly.

are labeled. Zhao *et al*. [172] used a very similar approach except that feature space is computed using LDA. Cai *et al*. [17] propose semi-supervised discriminant analysis algorithm that uses labeled data to infer the discriminative structure of the data while the intrinsic geometric structure of the data is inferred from both labeled and unlabeled samples.

4.2 Semi-supervised Person Recognition

To deal with the lack of labeled examples, we apply semi-supervised learning that leverages unlabeled examples for recognition. We will provide an example that demonstrates the motivation of our solution. We consider AR database [94] that consists of 100 subjects. Training set is created using a single image for each of the 100 subjects and rest of the images are used for testing. The highly successful face recognition algorithm Sparse Representation Classifier (SRC) [150] is used as a baseline. The first row of Figure 4.1 shows a query (Figure 4.1(a)) and the first 10 training samples (Figure 4.1(b)) belonging to 10 different subjects. SRC correctly predicts the identity of query whose training sample is quite similar to the query. However, with the same training set, the query in Figure 4.1(c) is not recognized correctly. This is due to the significant expression difference between the query and the corresponding training sample. Now, we augment the training set with one additional training example that exhibit expression variation for each subject as shown in Figure 4.1(d). We try to re-identify the query Figure 4.1(c) using SRC. With the large training set, SRC is able to correctly identify the label. While this is a very specific case, we note that the observations holds true across databases. In practical

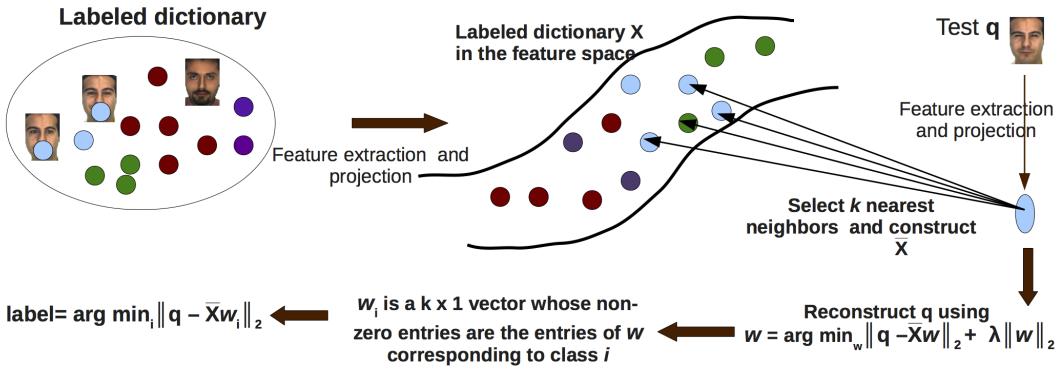


Figure 4.2 Overview of our proposed NNSC algorithm. Test image is represented as a linear combination of its nearest neighbors and assigned a label of the class that gives minimum reconstruction error.

situations, when there are limited training samples, supervised methods may not perform well as they are unable to account for all kinds of variations encountered in practical situations. Semi-supervised methods can be employed in such cases that use unlabeled samples to improve the performance of the recognition system. Our approach consists of *self-training* and *label propagation* algorithms. It is an iterative algorithm that hypothesizes the labels for the most confidently classified unlabeled samples using the label propagation algorithm. These confident samples are augmented to the training set. The algorithm is summarized in Algorithm 1.

4.2.1 Label Propagation for Person Recognition

We now look into the proposed label propagation algorithm that hypothesizes the labels of the unlabeled samples, and selects the most confident ones to augment to training set. Our setup consists of a training set $X = \{x_1, x_2, .., x_l, x_{l+1}..x_n\}$ where $x_i \in \mathbb{R}^d$. The samples with subscript $i = \{1, 2, .., l\}$ are the labeled samples that belong to class $c \in \{1, 2, .., C\}$ and remaining samples with subscript $i = \{l+1, l+2, .., n\}$ are unlabeled.

We construct an undirected graph $\langle V, E \rangle$ with similarity matrix W using both labeled and unlabeled samples. Each node in the graph corresponds to a face and the edges E represents similarities between them. Large edge weights w_{ij} indicate that corresponding faces are very similar. There are many metrics defined to measure the similarity between two points. A simple choice is k-nearest neighbor method to compute the weights where w_{ij}=1 if x_i is among the k-nearest neighbors of x_j or vice-versa and 0

otherwise. Another option for similarity measure is Gaussian function: w_{ij} = exp$(-||x_i - x_j||^2/2\sigma^2)$ where σ controls the spread of the Gaussian function.

For face recognition, the above mentioned similarity measures are not robust due to their sensitiveness to illumination, occlusion and expression variations. Representing a face as a *linear combination of training samples* is found to be very robust to such variations as observed in SRC [150]. Such representation also produces better weights that reveal the relations among different face samples. However, such representation over entire training set ignores the neighborhood information and may produce non-zero weights even for very dissimilar samples. Based on this observation, we propose a nearest neighbor based sparse coding (NNSC) that considers both locality and sparsity. In this representation, each sample is represented as a *linear combination of its nearest neighbors*. This is demonstrated in Figure 4.2. Finally, we use NNSC representation to construct the similarity weights of the graph.

Given a sample x_i, we obtain its NNSC representation as

$$\hat{w_i} = \arg\min_{w_i} ||x_i - B_i\, w_i||^2 + \lambda||w_i||_2 \;\text{ s.t }\; \forall_k\, w_{ik} \geqslant 0 \tag{4.1}$$

where the columns of B_i are the k-nearest neighbors, $\mathsf{N}(x_i)$ of x_i. We include the l_2-based regularization term as it is found to be discriminative [162] similar to l_1 regularizer. We also include an additional non-negative constraint to ensure that graph weights are greater than 0. λ denote the Lagrangian constant that controls the trade-off between the reconstruction error and the weights.

Once the representations are obtained for all the samples, we construct the similarity matrix $W \in \mathbb{R}^{n\times n}$ as:

$$W_{ij} = \begin{cases} \hat{w_i}(p), & \text{if } x_j \in \mathsf{N}(x_i) \\ 0, & \text{otherwise,} \end{cases}$$

where $i, j \in \{1, 2, .., n\}$ and $\hat{w_i}(p)$ denotes the p-th element of vector $\hat{w_i}$.

The obtained weights may not be symmetric i.e., $w_{ij} \neq w_{ji}$. We make the final weights symmetric with the operation: $w_{ij} = w_{ji} = (w_{ij}$+$w_{ji})/2$. We normalize the weights symmetrically as done in the spectral clustering to ensure convergence of label propagation algorithm as shown below.

$$L = D^{-1/2} W D^{-1/2} \;\text{ where }\; D_{ii} = \sum_j W_{ij} \tag{4.2}$$

Let $F \in \mathbb{R}^{n\times C}$ be a matrix from which the label y_i of a sample x_i, $\{i = 1, 2, .., n\}$ can be obtained as $y_i = \arg\max_j F_{ij}$, where $j = \{1, 2, .., C\}$. For the labeled samples x_i ($\{i = 1, 2, .., l\}$), we define Y_{ij} = 1 if $y_i = j$, and 0 otherwise. For unlabeled samples, $Y_{ij} = 0$, $\forall j$.

We assume that the class score of a sample can be computed as a linear combination of the scores of other samples, the weights being given from similarity weights w_i. We propagate the labels of the labeled samples to the unlabeled samples using the constructed graph weights. Using the label propagation framework, we let unlabeled samples receive some amount of label information from its neigbours and retain a part of its initial information at every iteration. The amount of information a sample receives is determined by its corresponding normalized weight.

We begin the iteration with $F(0) = Y$, and for any $t \geqslant 1$, the labeling matrix F is given by,

$$F(t+1) = \alpha L F(t) + (1-\alpha)Y, \tag{4.3}$$

where Y is the initial labeling of the samples and α is a parameter that decides the amount of information a sample receives at each iteration. It is well known from the label propagation literature [176] that the above iterative method converges to $F^* = (1-\alpha)(I-\alpha L)^{-1}Y$. Finally, the labels of unlabeled samples can be predicted using $y_i = \arg\max_j F^*_{ij}$.

4.2.2 Self-Training

Once the labels of unlabeled samples are predicted using the above algorithm, we retain the labels of most confident predictions and discard the rest. The most confident samples along with their predicted labels are added to the labeled set. To achieve this, we define the confidence measure based on the F^* scores. In the ideal case, each row of F^* contains a single non-zero component corresponding to the true class. The ratio of two largest components in a row in such case will be infinite. However in practice, the ratio will be large only when majority of the samples contributing to the reconstruction belong to a particular class. We use this ratio to measure how "confident" is the labeling decision after the convergence of label propagation. We consider the prediction as confident when the ratio of two largest labeling components of a sample i in $F^*_{ij} \; \forall \; j = \{1, 2, .., C\}$ exceeds a threshold. Another advantage of the confidence measure is that it allows to reject certain samples that are either outliers or out-of-database class images, which might otherwise reduce the performance on test images. Once the confident samples are added to the labeled set, label propagation is repeated with new labeled and unlabeled set and the self-training procedure is continued till convergence. The convergence can be based on either maximum number of iterations or desired number of labeled set are created. We will show in our experimental results, how this iterative self-training approach gives a significant improvement over a single stage

label propagation, especially when there are very few labeled samples and the dataset contains large intra-class variations. Our complete algorithm is summarized in Algorithm 1 and referred as ST-LP.

Algorithm 1 Our proposed algorithm (ST-LP) with self-training and label propagation.

1: **Input**: $X = [X_l\ X_u]$ where X_l and X_u denote label and unlabeled samples, respectively. C denote number of classes, number of self-training iterations K, count $= 0$.
2: $\bar{X_l} = X_l$ and $\bar{X_u} = X_u$
3: **while** count is less than K **do**
4: Apply Label Propagation algorithm using $\bar{X_l}$ and $\bar{X_u}$ as discussed in Sec 4.2.1.
5: Find the most confident set in $\bar{X_u}$ denoted as X_c
6: Augment the labeled set with confident unlabeled samples i.e. $\bar{X_l} = [\bar{X_l}\ X_c]$
7: $\bar{X_u} = \bar{X_u} \setminus X_c$
8: **end while**

4.2.3 Recognition of Target Examples

With the large training set, we perform classification of a novel test image using NNSC and reconstruction error classifier [150, 162]. For a test image $q \in \mathbb{R}^d$, its representation w over its k nearest neighbors, $B_q = [x_1, x_2, \ldots, x_k]$, is obtained using Eq. (4.1) but without any non-negative constraints. For each class *i*, we construct a function $\delta_i : \mathbb{R}^k \rightarrow \mathbb{R}^k$ which gets the coefficients associated with the *i*-th class in B_q.

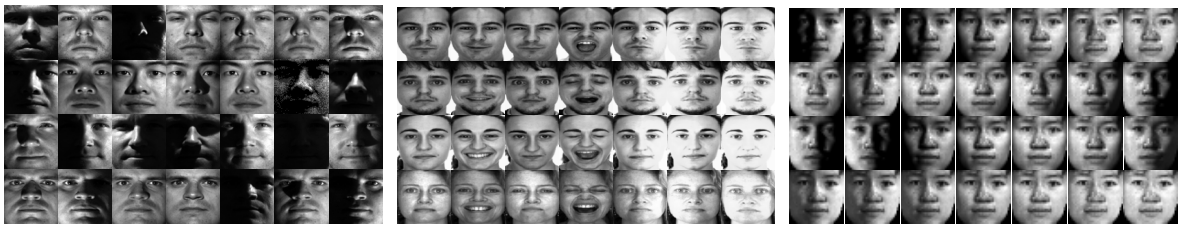
$$\delta_i^{(j)} = \hat{w}_j \quad \text{if } x_j \in \mathsf{N}(q);\ \ j = 1, 2, \ldots, k \tag{4.4}$$

where the non-zero entries of δ_i correspond to entries belonging to class *i* from w. The test image is then assigned the label of the class that minimizes the reconstruction error.

$$\text{label}(q) = \arg\min_i ||q - B_q\delta_i||_2 \tag{4.5}$$

4.3 Experiments and Results

We use Extended Yale B [72], AR [94] and CMU PIE [125] datasets to perform our experiments. Few examples from these datasets are shown in Fig. 4.3. We will use the Yale database to completely demonstrate the proposed algorithm and comparisons with previous methods are done using AR and CMU PIE datasets.



(a) (b) (c)

Figure 4.3 Examples from (a) Yale (b) AR and (c) CMU PIE databases. (a) and (c) have illumination and lighting variations while (b) has varying expression and illumination.

4.3.1 Semi-supervised Method

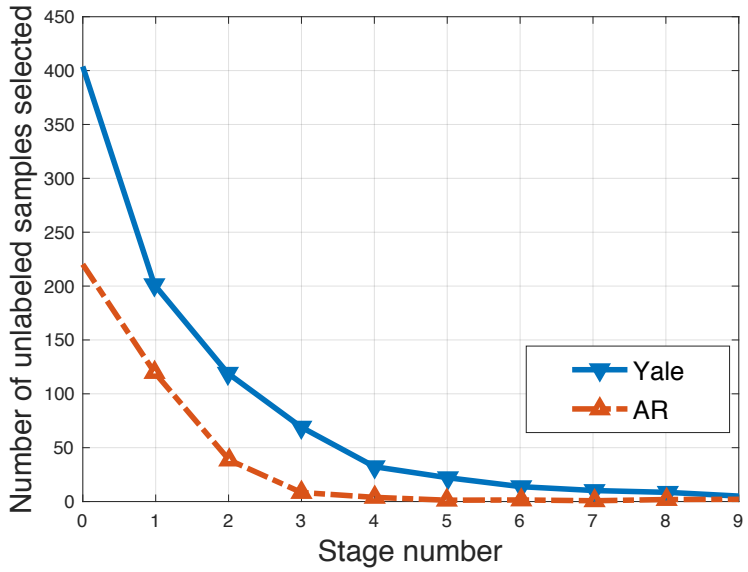
Extended Yale B database consists of 2414 frontal face images of 38 individuals captured under various lighting conditions. Each image is of size 192×168, which are resized to 80×80. Each subject consists of 64 images. We randomly select half of the images for training and other half for testing. As shown in Figure 4.3(a), the instances primarily exhibit illumination and lighting variations.

To create a semi-supervised setting, we keep the labels of only few training samples per class (3 to 24) and remove the labels for the rest of training samples. We select the maximum number of iterations K to 10 and the ratio of two largest labeling component F^*_{ij} (first and second largest labeling component) to $1 : 2.5$. We choose *k*, that indicates number of nearest neighbors in NNSC to 120. We use both labeled and unlabeled samples to find the feature space using PCA. We select the dimension of Eigenface to 504, $\alpha = 0.9$ and $\lambda = 0.01$.

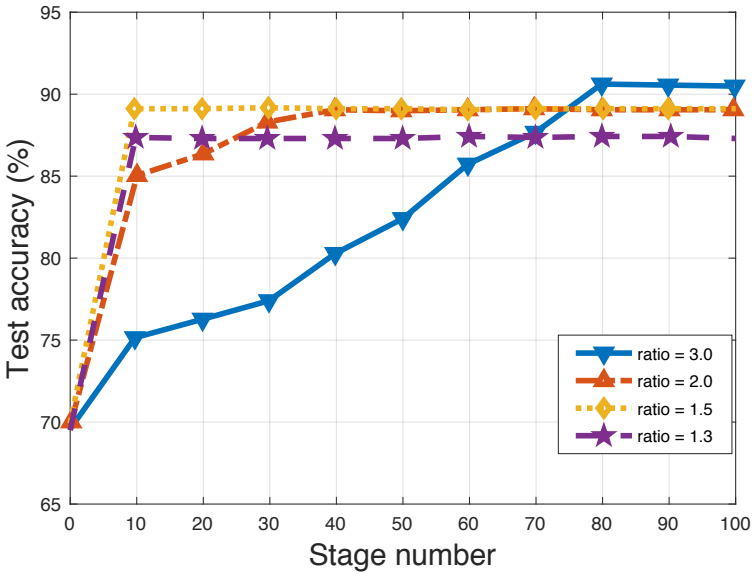
For different trials, we keep the labels of 3, 5, 10, 16 and 24 training samples in our experiments and labels of rest of the samples are removed. To measure the performance of our iterative approach, test accuracy is calculated after every iteration. Figure 4.5(a), shows the recognition rates of the test set after every stage for initial labeled samples 3, 5, 10, 16 and 24. It is clear from the figure that for every stage there is an improvement in the accuracy.

When the number of labeled samples considered are 16 and 24, gain in accuracy is not significant as there are enough labeled samples already available. The significant gain in accuracy can be observed for the trails with 3 and 5 labeled samples where there is almost 20% increase in the recognition rates that indicates the advantage of semi-supervised methods when there are limited training samples.

Figure 4.4(a) shows the number of unlabeled samples selected after every stage for Yale and AR database with the ratio of two largest labeling component $F^*_{ij} = 1.5$ and 3 labeled examples per subject.

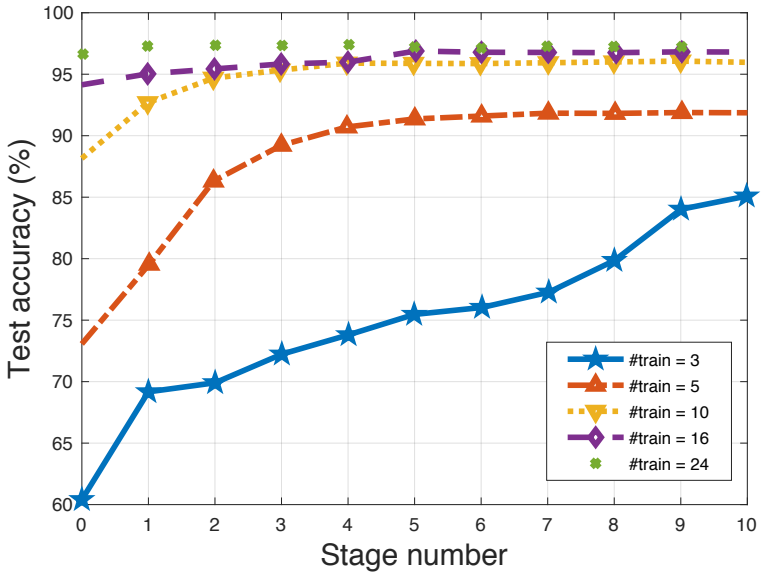


(a)

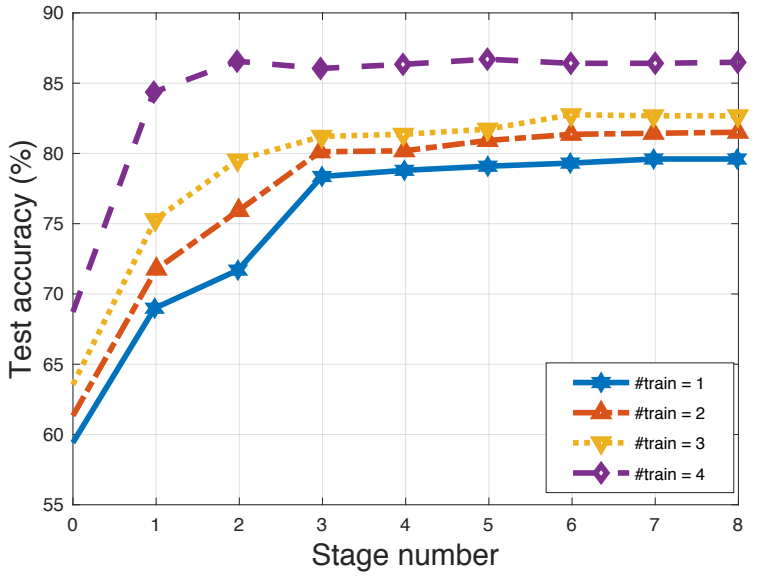


(b)

Figure 4.4 (a) **Number of unlabeled samples selected after every stage** on Yale and AR database for three labeled examples and threshold ratio=1.5. (b) **Effect of convergence and accuracy** for various values of ratio threshold on Yale database with three labeled samples.



(a)



(b)

Figure 4.5 (a) **Recognition rates on (a) Extended Yale B database and (b) AR database** for different iterations of our algorithm with different number of initial labeled samples.

Table 4.1 Recognition rates [%] of various methods on AR database for different number of labeled examples.

Method	1 Train	2 Train	3 Train
SRC [150]	54.4	61.2	65.4
CRC [162]	55.4	62.1	65.5
PCA self-training [113]	62.0	71.0	66.0
LDA self-training [172]	74.5	77.8	80.3
NNSC	59.4	61.6	66.0
ST-LP	**79.5**	**81.4**	**82.6**

It is clear that, large number of training samples are selected in the first few stages and hence larger gain during these stages. These highly confident samples selected in the first few stages will help in labeling the hard samples in the later stages.

The confidence threshold is an important parameter that decides the trade-off between performance and convergence. We conduct experiment with different threshold ratio and plot the results in Figure 4.4(b). A very high threshold takes long time to converge but reaches maximum performance. On the other hand, small thresholds converge much faster and produce a lower performance.

4.3.2 Comparison with Other Methods

AR database: We consider a subset of AR face dataset consisting of 50 male and 50 female subjects. For each subject there are 14 images with varying expressions and illuminations. Each image is of size 165×120. We convert the images to grayscale and resize to 80×80. Images are taken in two different sessions. We select seven images from session one for training and remaining seven images from session two for testing. Few images from this database are shown in Fig. 4.3(b). We reduce the dimension of the images to 504 using PCA. We select $\alpha = 0.9$ and $\lambda = 0.1$. For SRC, we select PCA dimension such that it maintains 75% over-completeness of the training set. We select the error tolerance, $e = 0.05$ for SRC and choose L1LS [64] l_1-regularized least squares solver to solve the minimization problem in SRC. We set the number of nearest neighbors k for NNSC to 100, the ratio of two largest labeling component F^*_{ij} to $1 : 1.5$ and maximum number of stages to 10. We conduct three trials with 1, 2 and 3 labeled samples. The recognition rates of the test set at various stages is shown in Fig. 4.5(b).

Table 4.2 Recognition rates on CMU-PIE database in (mean±std-dev%).

Method	Unlabeled set	Test set
Eigenface	25.3±1.7	25.3±1.6
Laplacianface	56.1±2.3	56.4±2.4
LapSVM [10]	56.5±1.6	56.9±2.6
LapRLS [10]	57.5±1.6	57.9±2.6
SDA [17]	59.0±2.0	59.5±2.7
LDA self-training [172]	84.5±9.5	71.3±6.5
SRC [150]	74.7±1.32	74.9±1.3
CRC [162]	74.9±1.41	75.1±1.32
NNSC	75.0±1.35	74.3±1.3
ST-LP	**92.1±1.3**	**92.3±1.5**

Table 4.1 compares the performance of our self-training label propagation (ST-LP) approach with other methods. It is clear from the table that, the performance of our algorithm is superior than previously proposed semi-supervised methods.

4.3.3 Single Training Sample Person Recognition

We now conduct the experiment with an extreme case in which a single training example is available for each subject. This setting is popularly referred as "one-shot learning" in the literature. We conduct our experiment on CMU PIE [125] dataset and compare the results with previously proposed supervised and semi-supervised algorithms. CMU PIE consists of 68 subjects with $41,368$ face images with varying illumination, pose, expression and lighting. As reported in [17, 172], we choose a subset of only frontal faces (C27) with only illumination and lighting variations that results in 43 images per subject. The images are cropped to 32×32. We used PCA to reduce the dimension of the image to 504 and k is set to 50, $\alpha = 0.9$ and $\lambda = 0.01$. For SRC, dimension is reduced to 50 to have an overcomplete dictionary. We set the maximum stages to 15 and the ratio of two largest labeling component F^*_{ij} to $1 : 1.5$. For any trial, 30 images are selected for training and remaining 13 are selected for test. Among the 30 training images, only one image is randomly selected and labeled, and remaining 29 samples remain unlabeled. The experiment is carried out 20 times and the results are averaged over 20 trials. The results in Table 4.2 show superiority of our proposed algorithm compared to other methods.

4.3.4 Effect of Parameters

Our algorithm consists of parameters k, λ and α. We observed that λ and α slightly affect the performance. From our experiments, we noticed the best results are obtained with $\alpha = 0.9$ and λ in the range $0.01 - 0.1$. We empirically selected the parameter k, number of nearest neighbors in NNSC and found that the algorithm is not very sensitive to k. A typical value in the range $100 - 200$ seemed to be robust and work well in practice.

4.4 Summary and Discussion

In this chapter, a semi-supervised solution using a novel nearest neighbor based linear representation is proposed for person recognition. The approach is particularly useful when labeled training data is limited. By exploiting the unlabeled data, it can better model the variations across instances, thereby providing better generalization. Our solution leverages unlabeled samples to improve the performance in a semi-supervised framework. We propose a robust approach based on self-training and label propagation algorithms that iteratively hypothesize the class labels of unlabeled samples and augments into the training set. For measuring the similarity between the samples, a novel similarity measure based on nearest neighbor sparse representation is introduced. The representation is faster and performs comparable to much expensive l_1-norm based sparse representations. Extensive experiments on classic face recognition datasets show the superiority of our solution compared to previous algorithms, including both supervised and semi-supervised methods.

The approach makes an assumption that there are sufficient unlabeled samples of subjects that needs to be identified are available. This may be restrictive in certain applications. In situations where it is not possible to generate unlabeled examples for in-class subjects, our approach can be used in conjunction with unsupervised dictionary learning techniques [107] to obtain the person representation and then incorporating into the graph framework.

The approach is generic and flexible. It can be used for recognizing other objects, and can be coupled with state-of-the-art deep features [116, 135, 104] that is the current research trend in face/person recognition community. We note that deep CNN features achieve near 100% performance on these constrained datasets and the additional performance improvement due to unlabeled samples becomes less obvious. For practical cases, however, we argue that the combination of deep features and our algorithm as a classifier should produce better results as we demonstrate in the next chapter.

Chapter 5

Improving Recognition with Domain Information

5.1 Introduction

In the previous chapter, we considered the issue of limited labeled samples for person recognition and a possible way to overcome it with unlabeled samples. In this chapter, we consider another issue that is relevant in practical deployment. We begin with the motivation of our problem.

Consider popular consumer and commercial applications such as photo-album tagging or person indexing in videos that require recognition of people in a *collection of images*. By a collection, we mean a set of images with multiple instances of a limited number of subjects. A straight forward strategy to follow in such scenarios is to apply off-the-shelf detectors and recognizers [69, 77, 116, 135] that are already proven to perform well on standard detection and recognition benchmarks (FDDB [56] or LFW [55]). There are two issues with this strategy. First, these algorithms are *generic*, aimed at recognition in any arbitrary image, and are trained on a large and diverse training set. The distribution of training instance may differ significantly with that of testing instances. This may happen due to different camera conditions, background, or dataset bias. This *domain-mismatch* may result in degrade the performance of algorithms. Second, these generic algorithms ignore the similarities among instances in the collection. For instance, a person may appear more than once in photo-albums. We hypothesize that techniques that exploit the similarities among instances in collections to make the predictions would lead to a better performance in such scenarios.

With this motivation, we consider the problem of person identification in a collection of images given a database of subjects, whose distribution differs significantly with the collection instances. Similar to previous chapter, we keep our attention on recognition with faces and highlight that the approach can be extendable to other kinds of person recognition.

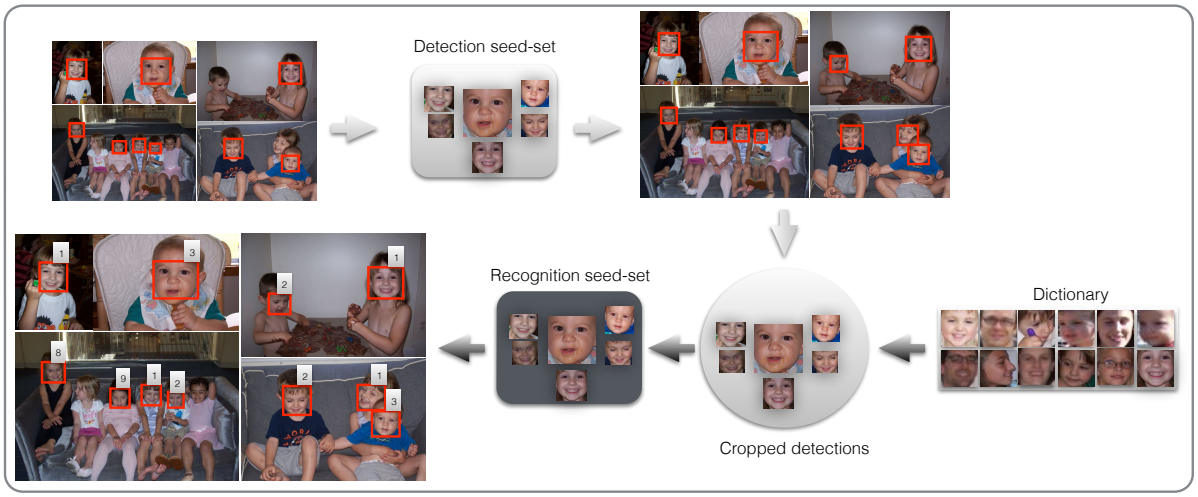


Figure 5.1 Overview of our person recognition approach in image collections. We detect and recognize faces in image collection in two stages. In the first stage, confident seed images are generated from the collection using off-the-shelf detection and recognition algorithms. In the second-stage, the obtained seed images are used to further improve the performance of both detection and recognition using the semi-supervised approaches.

We provide an overview of our approach in Figure 5.1. Our approach detects and recognizes faces in image collection in two stages. In the first stage, a set of seed instances are identified from the image collection using off-the-shelf face detection and recognition algorithms. In the second stage, the seed images obtained from the target domain are used to improve the performance of these algorithms in a semi-supervised framework. Our approach exploits the presence of large correlation among the faces in the collection and is suitable for offline applications.

For face detection, we use exemplar detector [69, 123] due to its superior performance and flexibility to retrain during semi-supervised learning. We select a few highly confident and diverse examples from the initial detection and clustering. The exemplar detector is then retrained on these seed examples to adapt it to the new domain. To recognize the detected faces, we follow a novel two-stage approach. We initially recognize the faces using a dictionary of labeled examples using off-the-shelf recognition algorithm and retain examples that are correctly labeled using a confidence measure. We then cast the recognition problem in a transductive semi-supervised framework treating the initially recognized faces from the collection as the labeled set and rest as the unlabeled set. We impose two constraints during propagation based on appearance and time space to exploit the relation among faces in different feature spaces.

Our approach achieves superior performance due to the following reasons. The appearance of the subjects that appear in multiple images of a collection is usually consistent. Our approach exploits this correlation among faces to propagate the labels from seed images to hard examples that are otherwise difficult to recognize. Also, unlike single stage recognition approaches [101, 136] that uses dictionaries with large number of subjects, our dictionaries in the second stage contain only the subjects present in the collection, thereby reducing the confusion during multi-class classification.

5.2 Related Work

Face detection is considered to be one of the classical computer vision problems can be roughly group into four categories: Adaboost/Boosting, Deformable parts based, Exemplar and Deep learning based models. A detailed description of these approaches are provided in chapter 2.

Semi-supervised face detection schemes are typically employed when the number of labeled examples are limited, or to adapt the detectors to the target domain. Lie *et al.* [75] trains an SVM classifier using top scoring positive and negative detections obtained using vanilla Viola-Jones. Similarly in [57], less confident detections in each target image are re-scored through a Gaussian regression process trained using top positive and negative detections. This approach is not practical for large image collections as each target image requires separate training. Sebe *et al.* [118] trains a detector using few labeled examples along with a large number of unlabeled samples. The approach may be less relevant today due to the availability of large scale face datasets.

Face recognition is another popular problem where the earlier focus was on designing better hand-crafted features and learning low dimensional subspaces (refer [171] for a detailed survey). One of the highly cited approaches in recent years is sparse representation classifier (SRC) [150] that represents the target image as a sparse linear combination of the dictionary elements. The approach that uses reconstruction error to predict the labels is robust to noise and occlusion. For the similar topic of face verification, metric learning [47, 126] is proposed to learn an embedding in which similar faces are closer and dissimilar faces are farther apart. More recently, CNN based approaches [104, 116, 135] that learn a hierarchy of low-level to mid-level features through an end-to-end training are producing impressive results on various face recognition and verification tasks. For more details, please refer to chapter 2.

Semi-supervised face recognition approaches [17, 113, 172] are proposed when there are only a few labeled training samples available. Roli *et al*. [113] first compute the Eigen face templates using labeled samples. The labeled set for each class is then augmented with unlabeled samples that are nearest to each of these class templates. The approach is repeated till a convergence is reached. Zhao *et al*. [172] propose a similar approach, however, using linear discriminant analysis (LDA). Semi supervised discriminant analysis [17] uses labeled data to infer the discriminative structure of the data, and both labeled and unlabeled data to learn the structure of the data in a graph based framework.

Face recognition in videos is another related topic that contains several works in three categories- key frame, temporal model, and image-set based approaches. Key-frame based approaches [169, 45] recognizes a set of selective frames in a track rather than all the frames. A majority voting scheme is used to label the entire video from the key frames. Temporal model based approaches [28] model the face correlation between consecutive frames in a video based on non-rigid facial expressions and head movements. Image-set based approaches [120] consider the face tracks as image sets and model the distributions of face images in each set and compare the similarity of distributions for recognition. A complete review of these approaches are given in [121]. There are few works that exploit additional cues in sitcom videos such as clothing and audio [29, 136], relations among subjects [16] for recognition. A semi-supervised scheme is employed in [16] that uses the weakly labeled data to align the subtitles and speaking face tracks. These approaches are applicable when audio and subtitle information are available.

5.3 Improving Detection in Image Collection by Adaption

Given a image collection, our objective is to detect as many faces as possible by adapting off-the-shelf detector to the instances in the collection. The overview of our detection approach is shown in Figure 5.2. We obtain initial detections from the collection using a generic detector trained on large database of images. We then select a few highly confident examples from the initial detections that are diverse and dissimilar through clustering. The detector is then retrained on these seed examples to adapt it to the new domain.

We choose the exemplar based detector [69, 123] due its high performance, simplicity, and flexibility it offers for adaption. The approach aligns well with our objective to improve performance using instances from the target domain. The exemplar detector uses retrieval framework for detection. A

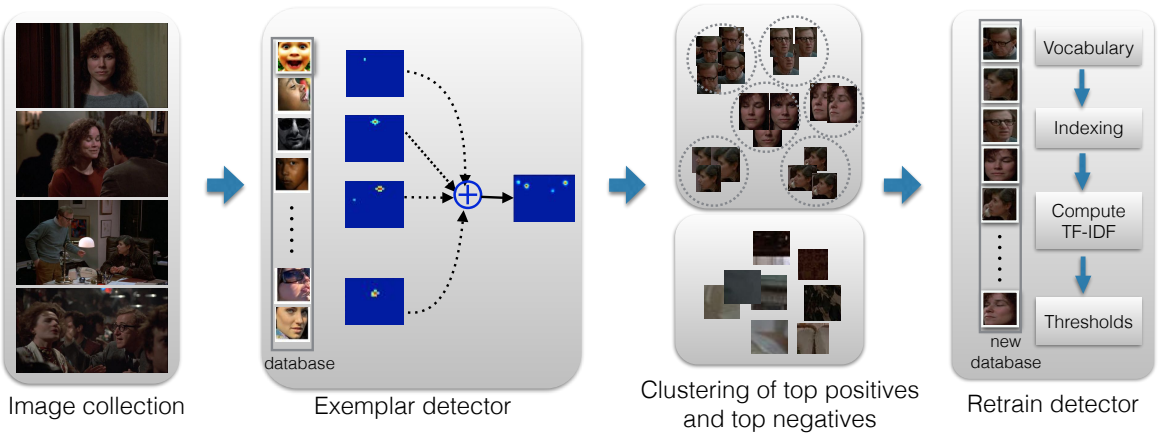


Figure 5.2 Overview of our detection approach for image collections. We apply off-the-shelf exemplar detector to get the initial detections. Top scoring positive and negative samples are selected and clustered to obtain a set of seed images. The exemplar detector is retrained by augmenting the seed images to exemplar database and re-computing the detector parameters.

large database of training exemplars that cover better facial variations is initially constructed and local features (dense SIFT) are extracted and quantized using a vocabulary. During testing, each exemplar generates voting maps using Hough voting scheme [36], which are then combined to locate the faces in the target image.

To achieve high recall, we initially apply the exemplar detector with very low threshold. Let d_i be the set of initial detections and s_i their corresponding scores. To obtain confident positive and negative instances, we define two different thresholds ω_1 and ω_2, which could be set either manually or estimated using a validation set. The detections with scores greater than ω_1 and less than ω_2 are considered as top positives and negatives, respectively.

$$\begin{aligned} P &= \{\, d_i, s_i > \omega_1,\ i = 1, 2, \ldots, n\} \\ N &= \{\, d_i, s_i < \omega_2,\ i = 1, 2, \ldots, n\} \end{aligned} \tag{5.1}$$

Once we obtain the sets P and N, we cluster the images using k-means and select τ detections as seed-images. This ensures that the obtained seed-images are diverse, which is essential for video collections containing near identical instances. If the diversity is not maintained, the most occurring exemplar instances dominate the voting process, severely degrading the performance. For each seed image, features are extracted and quantized, and augmented to the exemplar database. To retrain the exemplar detector, it is enough to update the inverse document frequencies of the visual words, unlike appearance based detectors [77, 96, 141] that require retraining of the models from scratch.

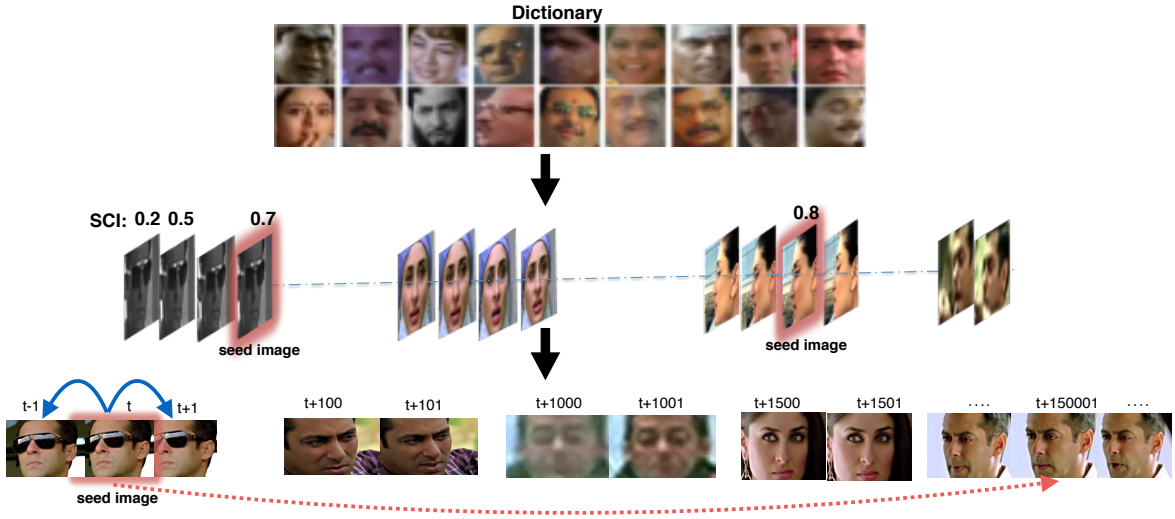


Figure 5.3 Overview of our domain adaptation approach for recognition in image collections. In the first stage, we label all the images in a collection using SRC and select only highly confident seed-images based on SCI. The labels of the seed-images are then propagated to remaining images in the collection by imposing the constraints in time and feature space.

5.4 Improving Recognition in Image Collection

We next recognize the detected faces using a labeled dictionary of subjects using a two-stage pipeline as shown in Figure 5.3. In the first stage referred as *seed-set selection*, the detections are recognized using off-the-shelf face recognition algorithm. We then identify the key seed images that are confidently recognized using a robust confidence measure. In the second stage referred as *propagation*, we propagate the labels from the seed images to the remaining unlabeled collection images incorporating various constraints.

Our recognition approach follows the intuition that certain faces in the collection may have similar appearance, pose to the faces in the given dictionary. If such faces from the collection are recognized correctly, it is possible to propagate their labels to the remaining unlabeled images as they belong to the same target domain. Our approach is thus applicable in scenarios where images are highly correlated. For instance, a video shot of zoom-in or zoom-out of a face in a fixed pose or video shot where there is a gradual change of pose of the subject as shown in Figure 5.4. If one or few images in such shots are recognized confidently in the first stage, the labels can be propagated to the remaining faces effectively through propagation.

Figure 5.4 A possible scenario that depict the effectiveness of our approach on a video collection with a head pose change. If a few images in such shots are recognized correctly in the first stage, their labels can be effectively propagated to the remaining images in the second stage.

5.4.1 Seed-Set Selection

We use Sparse Representation Classifier (SRC) [150] as a off-the-shelf face recognition algorithm for initial labeling as it is robust to noise and occlusion, and provides strong confidence measure based on reconstruction weights. Let $\mathbf{D} = [D_1, D_2, \ldots, D_M]$ be the dictionary of labeled examples and $\mathbf{X} = [X_1, X_2, \ldots, X_N]$ is the collection of face instances. SRC represents the target image as a *sparse* linear combination of dictionary faces.

$$\hat{\alpha_i} = \arg\min_{\alpha_i} ||X_i - D\ \alpha_i||_2 + \lambda||\alpha_i||_1, \tag{5.2}$$

where α_i is the representation of the sample X_i and λ is a Lagrangian constant that controls the trade-off between reconstruction error and sparsity. The label of X_i is obtained using the minimum reconstruction error criteria as

$$\text{label}(X_i) = \arg\min_{j} ||X_i - D_j\ \hat{\alpha_{ij}}||_2, \tag{5.3}$$

where D_j are the training samples belonging to class j and α_{ij} are the corresponding weights. Once the images in the collection are labeled, we select a few confident images using Sparsity Concentration Index (SCI) [150] defined as

$$SCI(i) = \frac{c \cdot \max_j ||\hat{\alpha_{ij}}||_1/||\hat{\alpha_i}||_1 - 1}{c - 1}, \tag{5.4}$$

where c denote the number of classes. This score indicates how well a target image is represented using the dictionary elements from a particular class. SCI score is very robust to noise and occlusion and serves as a strong factor to measure the recognition confidence. A high score indicates that the target image is represented mostly from the single class and a very low score close to zero indicates contribution from all the classes. We consider this SCI score as a confidence measure to retain the label of a face.

5.4.2 Propagation of Seed Images

Once confident seed images are selected, their labels are propagated to the remaining unlabeled collection images using a label propagation framework [176]. Since the collection contain correlated images, an image can be recognized using its appearance and its similarity with other images. This can be achieved in a graph based framework where the image similarities are used to predict their labels. Recognition using seed images in the second stage reduces the domain mismatch between the dictionary and collection, and also the confusion rate when the dictionary has large number of subjects compared to the collection.

We reconsider $\mathbf{X} = [X_1, X_2, \ldots, X_N] = [X^l\ X^u]$ without loss of generality. Here, X_l denote the labeled seed images obtained in the first stage and X_u denote the remaining unlabeled images in the collection. Let, $F \in \mathbb{R}^{N \times c}$ denote a non-negative labeling matrix where the i-th row of F is the labeling score of image X_i for each class. Let $Y \in \mathbb{R}^{N \times c}$ denote the initial labeling matrix. For the seed image i belonging to class j, we define $Y_{ij} = 1$, and 0 otherwise. For all the remaining unlabeled images, we assign a zero vector, i.e., $Y_{ij} = 0$, $\forall j$.

Given X, we construct an undirected graph $\langle V, E \rangle$ using both labeled and unlabeled images. Each node in the graph represents an image and the edges E represent the similarities between images. Larger the edge weight, greater is the similarity between images. The design of similarity measure usually depends on the task at hand. For album collection, we consider appearance similarity and for videos, we consider appearance and temporal similarities.

5.4.2.1 Temporal Similarity

Let t_i and t_j be the absolute numbers that denote i-th and j-th frames, respectively. Also let $\upsilon_i = (\upsilon_i^{x_1}, \upsilon_i^{y_1}, \upsilon_i^{x_2}, \upsilon_i^{y_2})$ and $\upsilon_j = (\upsilon_j^{x_1}, \upsilon_j^{y_1}, \upsilon_j^{x_2}, \upsilon_j^{y_2})$ denote the x and y co-ordinates of top-left and bottom-right corners of the rectangles representing i-th and j-th face in the collection, respectively. We define temporal similarity as follows,

$$W_{ij}^t = exp\left(\frac{-(t_i - t_j)\chi_{ij}}{2\sigma_t^2}\right),$$

where σ_t controls the spread of the Gaussian function. χ_{ij} is defined as the absolute sum of the differences of the i-th and the j-th image co-ordinates.

$$\chi_{ij} = \sum |\upsilon_i - \upsilon_j|.$$

A large value is assigned for W^t_{ij} for pair of faces if they appear in subsequent frames and at similar locations. With above similarity, we define the temporal constraint as

$$\sum_{i,j} W^t_{ij} ||F_i - F_j||^2. \tag{5.5}$$

During propagation, this constraint ensures that the images that are closer in temporal space have similar labels.

5.4.2.2 Appearance Similarity

Images that are similar in appearance space (SIFT, HOG or CNN) should have large edge weights between them. While weights based on k-nearest neighbor and Gaussian function are commonly used in literature, they are not robust to facial illumination and expression variations. For this reason, we employ a new scheme to compute the weights modifying the SRC representation. We represent each image as a linear combination of its nearest neighbors to preserve the locality and impose non-zero constraints on the weights. Our approach encourages the creation of edges only with similar samples, which is essential for obtaining good performance.

$$\hat{w_i} = \arg\min_{w_{ik}} ||X_i - \Sigma_{k:X_k \in \mathcal{N}(X_i)} X_k w_{ik}||_2 + \beta ||w_i||_2 \quad \text{s.t} \quad \forall_k,\ w_{ik} \geqslant 0, \tag{5.6}$$

where $\mathcal{N}(X_i)$ denotes the k neighboring samples of X_i, and β is a Lagrangian constant that controls the trade-off between two terms. Appearance weight matrix $W^a \in \mathbb{R}^{N \times N}$ is then constructed as:

$$W^a_{ij} = \begin{cases} \hat{w_i}(k), & \text{if } X_j \in \mathcal{N}(X_i) \\ 0, & \text{otherwise} \end{cases},$$

$\hat{w_i}(k)$ denotes the k-th element of vector $\hat{w_i}$ corresponding to k-th neighbor. Weights obtained by this method may not be symmetric, i.e, $W^a_{ij} \neq W^a_{ji}$. To make the weights symmetric, we perform the below operation

$$W^a_{ij} = W^a_{ji} = \frac{W^a_{ij} + W^a_{ji}}{2}.$$

The appearance constraint is finally incorporated into the label propagation framework as

$$\sum_{i,j} W^a_{ij} ||F_i - F_j||^2. \tag{5.7}$$

This constraint ensures that images that are highly similar in appearance space should belong to the same class.

5.4.2.3 Propagation

We finally incorporate appearance (Eqn 5.7) and temporal (Eqn 5.5) constraints into label propagation formulation [176] to propagate the labels from labeled seed images to unlabeled images as follows.

$$Q(F) = \arg\min_{F} \frac{\gamma_1}{2} \sum_{i,j}^{N} W_{ij}^{t} ||F_i - F_j||^2 + \frac{\gamma_2}{2} \sum_{i,j}^{N} W_{ij}^{a} ||F_i - F_j||^2 + \gamma_3 \sum_{i}^{N} ||F_i - Y_i||^2. \tag{5.8}$$

If the first and second terms ensure that the images that are similar based on appearance and temporal weights have similar labels, third term retains the labels of the confident seed images. γ_i are the parameters that control the trade-off between these three terms.

Let, $D_{ii}^{t} = \sum_j W_{ij}^{t}$ and $D_{ii}^{a} = \sum_j W_{ij}^{a}$ be the diagonal and symmetric matrices whose entries are row sums of W_{ij}^{t} and W_{ij}^{a}, respectively. $L^t = D^t - W^t$ and $L^a = D^a - W^a$ are the Laplacian matrices that are symmetric and positive semi-definite, defined on temporal and appearance similarities, respectively. Following [9], we can rewrite the first term in Eqn 5.8 as

$$\sum_{i,j}^{N} W_{ij}^{t} ||F_i - F_j||^2 = \sum_{i,j}^{N} (F_i^2 + F_j^2 - 2F_iF_j) W_{ij}^{t} \tag{5.9}$$

$$= \sum_{i}^{N} F_i^2 D_{ii}^{t} + \sum_{j}^{N} F_j^2 D_{jj}^{t} - 2 \sum_{i,j}^{N} F_i F_j W_{ij}^{t} \tag{5.10}$$

$$= 2\bar{F} L^t F \tag{5.11}$$

where $\bar{F}$ is the matrix transpose of F. Similarly, second term is equivalent to $2\bar{F}L^aF$. Thus, Eqn 5.8 can be rewritten as

$$Q(F) = \arg\min_{F} \gamma_1 \bar{F} L^t F + \gamma_2 \bar{F} L^a F + \gamma_3 ||F - Y||^2 \tag{5.12}$$

Differentiating $Q(F)$ with respect to F and equating to 0, we get

$$2\gamma_1 L^t F + 2\gamma_2 L^a F + 2\gamma_3 (F - Y) = 0 \tag{5.13}$$

$$F^* = \gamma_3 (\gamma_1 L^t + \gamma_2 L^a + \gamma_3)^{-1} Y \tag{5.14}$$

The final identity y_i of the image X_i can be obtained from F^* as $y_i = \arg\max_j F_{ij}^*$, where $j = \{1, 2, \ldots, c\}$.

5.4.3 Rejection of Unknown Faces

The candidate boxes obtained from the detection stage may contain false positives or unknown faces that are not present in the dictionary. This is particularly common in video collections. The algorithm

should be able to accept or reject the labels of such candidates after propagation. For this purpose, we define a confidence measure known as *label dominance score* (LDS) that measures the contribution of each class during the reconstruction of the sample. This is simply defined as the ratio of two largest class scores of the sample.

$$\text{LDS (i)} = \frac{F_{ij}}{\arg\max\limits_{k,\, k \neq j} F_{ik}} \text{ where } j = \arg\max_{j} F_{ij} \tag{5.15}$$

Intuitively, when an image has large edge weights with images belonging to a particular class, the scoring vector F_i will have a high score for that particular class. When there is such clear dominance of one class, LDS will be high and we consider the labeling as confident and retain its final label.

5.5 Experiments and Results

5.5.1 Datasets and Setup

We evaluate our approach on several image and video collections whose instances exhibit pose, illumination, view-point variations that are seen in the real world. Few images from these datasets are shown in Figure 5.5.

G-album [37] is a small collection of family photos captured at various indoor and outdoor events. It consists of 589 images containing 931 faces belonging to 32 subjects. For each subject, we consider a maximum of 10 images to create a labeled dictionary and the rest as unlabeled. This setting is similar to automatic photo-tagging of albums in commercial softwares where users provide the initial labeled set by tagging a few images.

People in Photo Albums (PIPA) [165] is a large collection of $1{,}438$ user-uploaded albums collected from Flickr. The dataset consists of $37{,}107$ photos with $63{,}188$ head instances belonging to $2{,}356$ identities. Similar to G-album setting, we consider a maximum of 10 images to create a labeled dictionary and the rest as unlabeled.

Movie Trailer Face Dataset [101] consists of $4{,}485$ face tracks from 101 movie trailers released in the year 2010. These trailers are collected from YouTube and contain the celebrities presented in the PubFig Dataset [67] along with additional 10 actors. The labeled dictionary consists of $34{,}522$ images (PubFigs + 10 additional actors) with each actor having a maximum of 200 images. Since the dataset provides the raw descriptors based on the combination of LBP, HOG, and Gabor features), we show only the recognition results.

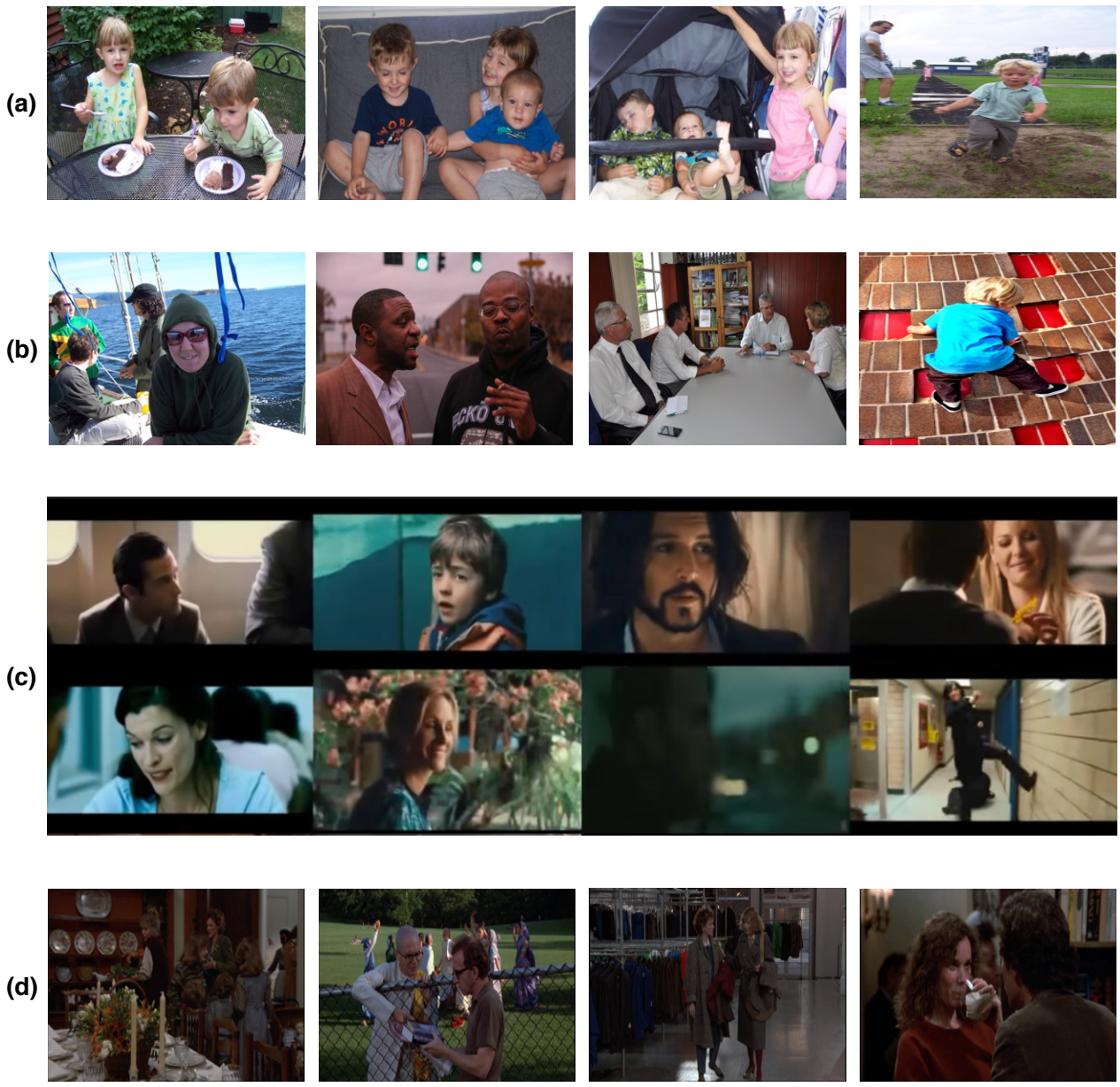


Figure 5.5 Image collection datasets: Few example images from (a) G-album (b) PIPA (c) Movie Trailer and (d) Hannah movie datasets. These image exhibit large variations in illumination, pose, viewpoint, occlusion, resolution, etc. and offer very challenging recognition scenario.

Table 5.1 Recognition rates [%] of various methods on G-album for different number of labeled examples.

Method	1 Train	2 Train	5 Train	10 Train	20 Train
KNN	56.12	65.85	74.15	77.26	80.43
SVM	57.45	64.40	78.09	80.82	85.42
SRC [150]	56.45	68.39	77.86	79.56	86.82
CRC [162]	56.27	68.28	77.62	75.33	85.87
Our approach	**60.25**	**71.27**	**82.52**	**84.52**	**87.18**

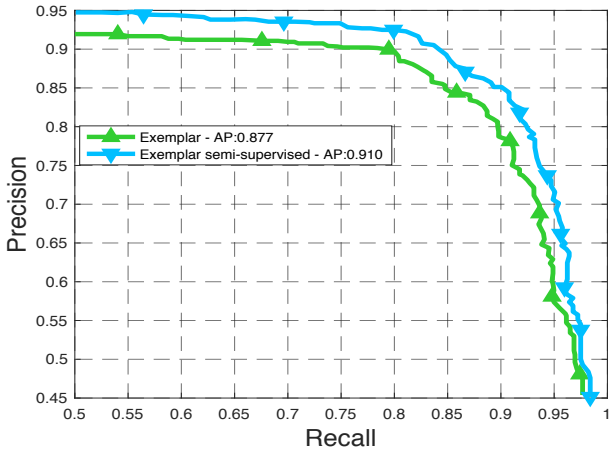
Hannah Movie Dataset [102] consists of face annotations for the entire movie *Hannah and Her Sisters*. The dataset has 153,833 frames with 202,178 face bounding boxes and 254 different labels of 41 named, 186 unknown characters, and 15 miscellaneous crowd regions. We create the labeled dictionary from IMDB photos[1]. For each named character, we collect photos from actor's profile in IMDB and annotate the face region. Of the 41 named characters, only prominent actors had profiles in IMDB. The IMDB dictionary consists of 2,385 images belonging to 26 prominent actors appeared in the movie. There are a total of 159,458 instances belonging to these 26 actors in the movie. There is a significant age variation and domain shift between the dictionary and movie instances since the Hannah instances are created from a particular year (1986) while IMDB photos are captured over a long period of time.

5.5.2 Results

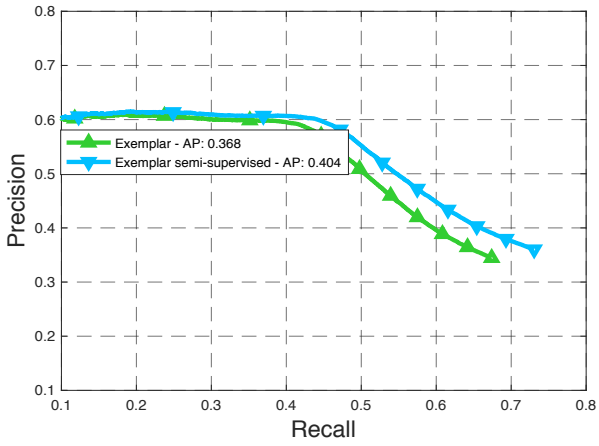
We use the publicly available implementation of [69] for exemplar face detection. We set the upper and lower thresholds ω_1 and ω_2 to 80-th and 20-th percentile of the detection scores to select the top 20% positives and negatives, respectively. For video collections, we further cluster the detections to select 300 diverse images. The obtained detections are resized and followed through the feature indexing pipeline as the original algorithm. The term frequencies and inverse document frequencies are updated based on seed images. The detection results in Figure 5.6 shows that performance improvement obtained using our approach.

We extract deep CNN features using VGG model [104] trained on 2M faces of 2,622 identities using `VGG-16` architecture. We then reduce the dimensionality of the deep features to 300 using PCA. For

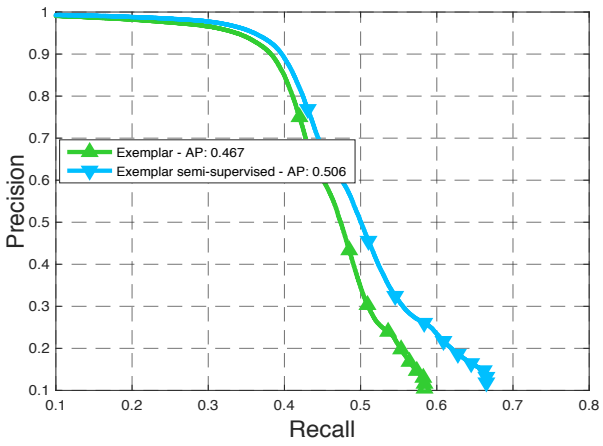
[1] Hannah and Her Sisters (IMDB): `http://www.imdb.com/title/tt0091167/`



(a) G-album

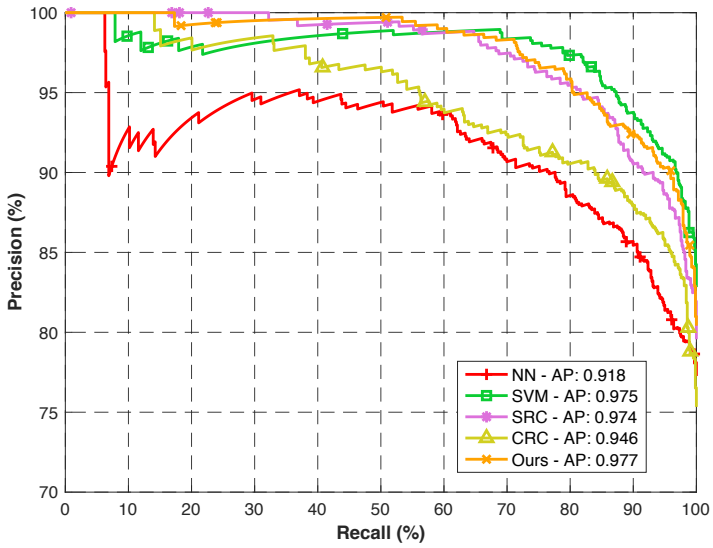


(b) PIPA

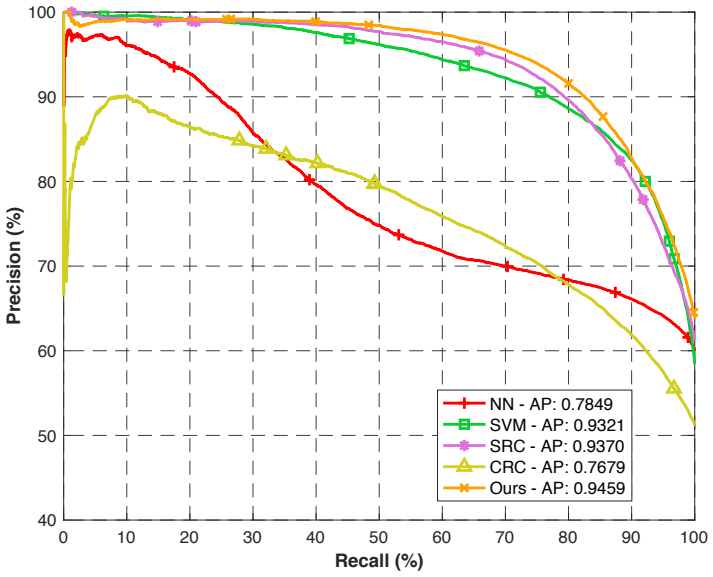


(c) Hannah

Figure 5.6 Precision-recall curves of *baseline* and *adapted exemplar detection* on various image collections. We notice a considerable improvement in the performance with our proposed approach that involves a simple augmentation of seed images into the exemplar database.

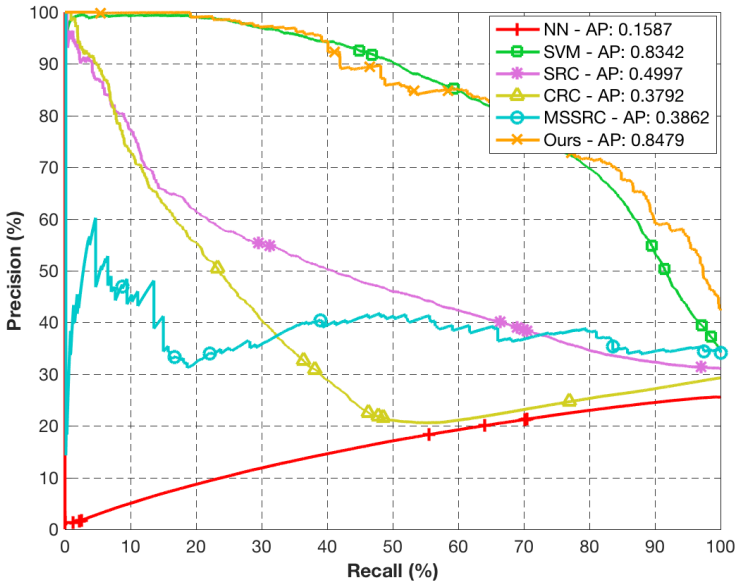


(a) G-album

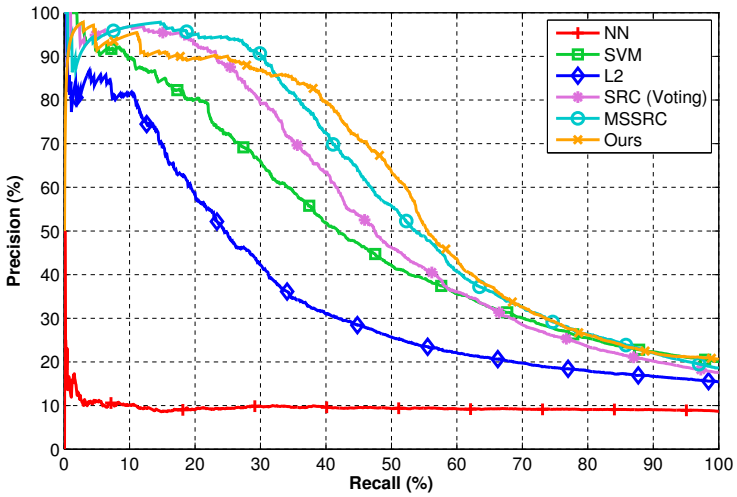


(b) PIPA

Figure 5.7 Precision-recall curves of *recognition* algorithms on album collections (a) G-album and (b) PIPA. Our confidence score based on LDS is more robust and performs better than all the approaches in rejecting unknown samples.



(a) Hannah



(b) Movie Trailer

Figure 5.8 Precision-recall curves of *recognition* algorithms on video collections of (a) Hannah and (b) Movie Trailer datasets. LDS score is robust in rejecting unknown samples.

Table 5.2 Recognition performance (%) of various methods in terms of accuracy and average precision (AP) on album collections.

Method	G-album		PIPA	
	Accuracy	AP	Accuracy	AP
1-NN	77.26	91.69	56.02	78.49
SVM	80.82	97.57	38.15	93.21
CRC [162]	75.33	94.58	51.27	76.79
SRC [150]	79.56	97.41	58.84	93.70
MSSRC [101]	79.56	97.41	58.84	93.70
Our approach	**84.52**	**97.72**	**63.80**	**94.59**

Table 5.3 Recognition performance (%) of various methods in terms of accuracy and average precision (AP) on video collections.

Method	Hannah		Movie Trailer	
	Accuracy	AP	Accuracy	AP
1-NN	32.71	15.87	23.60	9.53
SVM	44.82	83.42	54.68	50.06
CRC [162]	37.49	37.92	41.93	36.33
SRC [150]	39.76	49.97	47.78	54.33
MSSRC [101]	43.74	38.62	50.52	58.69
Our approach	**54.19**	**84.79**	**55.98**	**59.34**

generating the seed images during recognition, we set the error tolerance $\lambda = 0.05$ and retain the labels of top 25% of images based on SCI. We set various parameters using cross-validation on held-out set. We set the number of neighbors k for computing appearance weights to 120, $\gamma_2 = 0.3$ and $\gamma_3 = 0.7$. Whenever collection has temporal information, we set $\gamma_1 = 1$, otherwise 0.

We show the the recognition performance of various approaches on image and video collections in Table 5.2 and Table 5.3, respectively. Our approach that exploits the correlation among the instances outperform other approaches that recognize instances independently. Note that, for image collections without tracks, MSSRC is same as SRC. We also show the average precision that measures the ability to reject unknown instances. We use SCI as a confidence measure for SRC, CRC, MSSRC and LLC algorithms, L2 Euclidean distance for k-NN and probability scores for SVM. We show the precision-recall (PR) curves of various methods in Figure 5.7 and Figure 5.8 for album and video collections, respectively. It is clear from table that our confidence measure based on LDS is robust in rejecting

Figure 5.9 Performance improvement after adapting the exemplar detectors to the image collection from (a) G-album and (b) Hannah datasets. The original detections are shown in blue and new detections are shown in red. Notice how the faces that were by missed the original detector are detected after adaption using images from the seed-set.

unknown instances. The recognition rates of various approaches for different number of labeled training samples are shown in Table 5.1 for G-album dataset. The performance improvement is larger when there are limited labeled examples and it becomes closer to other approaches as the labeled examples are increased. Finally, we show the performance of our approach for varying percentage of seed images retained in Figure 5.10. The performance improvement is significant when we retain $20-30\%$ of seed images and becomes closer to the performance of first stage with increasing proportion of seed images. We show few qualitative results of our detection and recognition approach in Figure 5.9 and Figure 5.11, respectively.

Computational complexity: Our approach is slightly expensive compared to other approaches due to the two-stage process. With the feature extraction common across stages, the additional complexity is due to voting map generation during detection and $<W^t, W^a, F>$ computation during recognition in the second stage. The detection and recognition steps take ~ 1.5 and ~ 2 times more than the single stage approach, respectively. Our approach brings substantial performance improvements with an additional cost that is affordable for offline applications.

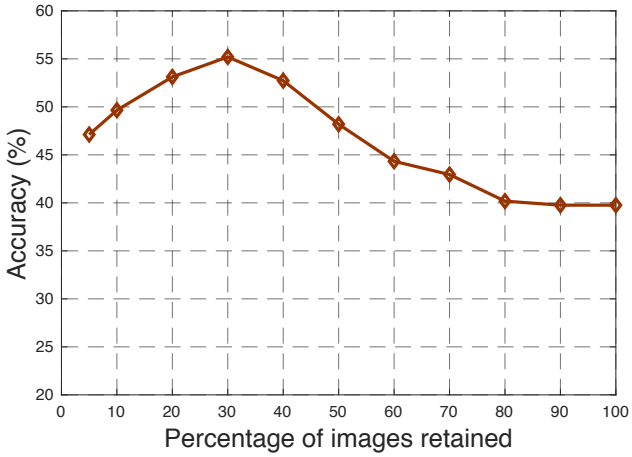


Figure 5.10 Recognition performance for varying percentage of seed-images selected in the first stage using Hannah dataset. The performance improvement is significant when $20-30\%$ of seed images are selected, and becomes less effective when large proportion of seed images are retained.

5.6 Discussion

A novel two-stage person detection and recognition approaches that adapt to the target domain of image collection is proposed. The main idea is to generate a confident set of seed images from the target using off-the-shelf algorithms and then exploit them to predict the labels of remaining images in the target domain. The approach can be used to recognize people in challenging scenarios in which the appearance and imaging conditions of target images is significantly different compared to gallery images. Our work is related to the literature of domain adaption [105] which is a well studied topic in computer vision. In domain adaptation, labeled data from source domain is used to train models to perform well on a different yet related target domain.

A particular recognition scenario with image collections is considered in this work. Our proposed work considers collective recognition of large number of target images to obtain the performance improvement. As a result, it may become less relevant for applications where individual instances needs to be identified. This is one of the primary differences of our work compared to the latest literature where the focus is to obtain improved performance on standard benchmarks such as `LFW` [55] or `PIPA` [165].

Our approach is complementary to current research focus where a predominant number of research papers in the recent years are focused on learning deep CNN representations. While we have used VGG features, it can be used with other deep features such as DeepFace [135], FaceNet [116], *etc*. We have considered appearance and temporal constraints in our work. The proposed framework is flexible

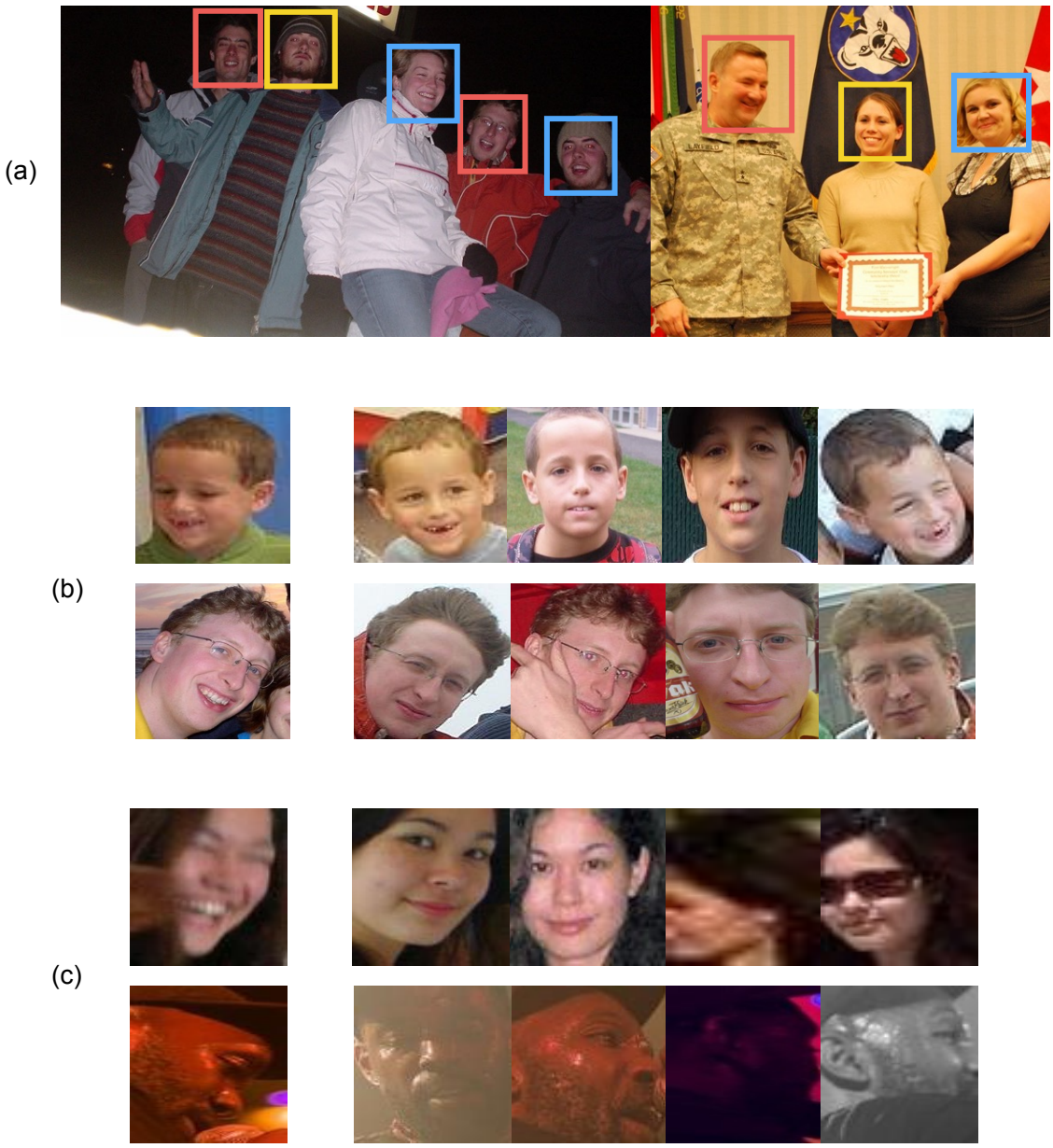


Figure 5.11 Qualitative results on PIPA: (a) shows the comparisons of our approach with single stage approach of SRC. Red boxes show the success case of our approach and failure case of SRC. Blue (and yellow) boxes indicate instances that are (not) correctly predicted by both approaches. (b) and (c) show the success and failure cases of our approach, using the test instance and its top-4 instances with non-zero appearance weights w^a.

and allows to include other constraints such as relationships between instances [74, 13], meta-data (subtitles, clothing, location, background, *etc*.) [7, 157] and multiple appearance cues (hair, upper body, scene) [62]. Finally, sparse representation classifier is considered for generating the seed images during recognition. It is possible to consider other recognition algorithms provided the output class scores are robust for measuring the confidence predictions.

5.7 Summary

We present an approach for people recognition in image collections that contain multiple instances for each subject. For this problem, we follow domain adaption strategy using a novel two-stage approach for both detection and recognition tasks. We first generate a set of seed images using off-the-shelf detection and recognition algorithms. The generated seed images are then used to improve the performance of detection and recognition by adapting them to the target image collection. We adapt the exemplar based detector to obtain improved detections and use label propagation based framework for improving the recognition. Our approach exploits the similarities among the images in the collection to improve the performance. Experiments on various real-world image collections suggest that our method performs better than previous approaches that identify the images independently.

Chapter 6

Person Recognition with Multiple Body Cues

Figure 6.1 *Beyond Frontal Faces for Recognition:* We often rely on complementary cues such as hair, accessories, clothing for recognition. Recognizing people with multiple body cues is the focus of our ongoing work.

6.1 Introduction

In the previous chapters, we focused on people recognition using faces. However, in many practical scenarios such as social media photos, surveillance, sport or entertainment videos, faces are not completely visible in images, and may be occluded, be of low resolution, facing away from the camera or even cropped from the view. As a result, we turn our attention towards such real-world scenarios where it becomes necessary to look for additional cues for recognition.

We will start with a motivating example in Figure 6.1. The images shown in Figure 6.1 are popular personalities whose face is not completely visible. They are either occluded, facing away from the camera, or have extreme pose. Yet, we can easily identify the public personalities present in these pictures based on our knowledge about their dressing, hair style, profession and so on. We could identify for instance Mandela from his unique hair style, Pope and Mother Teresa from their unique clothing, and Roger Federer, probably through a combination of face, hair and accessories. In addition to facial cue,

Figure 6.2 *Change of appearance with pose:* Each row shows the appearance of the same person in different poses. Note that the distinguishing features of a person that help in classification are different in each pose.

we relied on other body cues to support or strengthen our decision making process about the subject. If recognition systems rely entirely on faces, they would perform poorly in such situations. For instance, the very successful face recognition engine from Facebook "Deep Face" [135] which achieved close to 100% performance on LFW [55], could achieve only 46.66% on PIPA dataset [165] where many instances do not have visible frontal faces. With this motivation, we focus on new line of work for practical person recognition where multiple body regions are used to achieve improved recognition performance.

Another major challenge in person recognition or any fine-grained object recognition in general is the pose and alignment of different object parts. The appearance of the same object changes drastically with different poses and view-points (rows of Figure 6.2) causing a serious challenge for recognition. One way to overcome this problem is through pose normalization, where objects in different poses and view-points are transformed to a canonical pose [12, 21, 49, 135, 158, 178]. Another popular strategy is to model the appearance of objects in individual poses by learning view-specific representations [3, 61, 95, 164].

In this chapter, we aim to learn pose-aware representations for person recognition. While it is straight-forward to align objects such as faces, it is harder to align human body parts that exhibit large variations. Hence we design view-specific models to obtain pose-aware representations. We partition the space of human pose into finite clusters (columns of Figure 6.2) where each cluster contain samples in a particular body orientation or view-point. We then learn multi-region convolutional neural network

(convnet) representations for each view-point. However, unlike previous approaches that train a convnet for each body region, we jointly optimize the network over multiple body regions with a single identification loss. This provides additional flexibility to the network to make predictions based on a few informative body regions. This is in contrast to separate training that strictly enforces correct predictions from each body region. The joint network also converges much faster during training and simplifies the feature extraction process during testing. During testing, we obtain the identity predictions of a sample through a linear combination of classifier scores, each of which is trained using a pose-specific representation. The weights for combining the classifiers are obtained by a pose estimator that computes the likelihood of each view.

Our approach overcomes some of the limitations of the previously proposed approaches, `PIPER` [165] and `naeil` [62]. Although poselet-based representation of `PIPER` normalizes the pose; individual poselet patches [15] by themselves are not discriminative enough for recognition tasks, and under-perform compared to fixed body regions such as head and upper body. On the other hand, `naeil` learns a pose-agnostic representation using more informative body regions. Our framework is able to combine the best of both approaches by generating pose-specifc representations based on discriminative body regions, which are combined using pose-aware weights.

Our other major contribution is the rigorous evaluation of person recognition by creating two additional benchmarks. Current approaches [62, 74, 165] have solely focused on the photo albums scenario, reporting primarily on PIPA dataset [165]. However, this setting is very limited due to the similar appearance of people in albums, clothing and scene cues. To create a more challenging evaluation, we consider three different scenarios of photo albums, movies and sports and show significant performance improvements with our proposed approach.

6.2 Related Work

Person recognition has been attempted in multiple settings, each assuming the availability of specific types of information regarding the subjects to be recognized.

Face recognition is by far the most widely studied form of person recognition. The area has witnessed great progress with several techniques proposed to solve the problem, varying from hand-crafted feature design [6, 106, 149], metric learning [47, 126], sparse representations [150, 166] to state-of-the-art deep representations [104, 116, 135].

Person re-identification is the task of matching pedestrians captured in non-overlapping camera views; a primary requirement in video-surveillance applications. Most popular existing works employ metric learning [44, 53, 159] using hand-crafted [31, 86, 91] or data-driven [5, 80, 170] features to achieve invariance with respect to view-point, pose and photometric transformations. The approaches in [139, 159] also optimized a joint architecture with siamese loss on non-overlapping body regions for re-identification.

Pose normalization and **multi-view representation** are the two common approaches in dealing with object pose variations. Frontalization [49, 135, 158, 178] is a pose normalization scheme used commonly in face recognition, where faces in arbitrary poses are transformed to a canonical pose before recogniton. Pose-normalization is also applied to the similar problem of fine-grained bird classification [12, 163]. Unlike rigid objects such as faces, it is difficult to align human body parts due to large deformations. Hence we follow a multi-view representation approach where the objects are modeled independently in different views. This has also been employed in face recognition [3, 95], where training faces are grouped into different poses and pose-aware CNN representations are learnt for each group.

Person recognition with multiple body cues is the problem of interest in this work. We make direct comparisons with the recent efforts that use multiple body cues: `PIPER` [165], `naeil` [62] and Li *et al.* [74]. `PIPER` uses a complex pipeline with 109 classifiers, each predicting identities based on different body part representations. These include one representation based on Deep Face [135] architecture trained on millions of images, one AlexNet [66] trained on the full body and 107 AlexNets trained on poselet patches, the latter two using PIPA [165] trainset. On the other hand, `naeil` is based on fixed body regions such as face, head and body along with scene and human attribute cues trained using four different datasets, namely PIPA, CASIA [160], CACD [19] and PETA [25]. While poselets (used in `PIPER`) normalizes the pose, they are less discriminative compared to fixed body regions employed by `naeil`. We combine the strengths of both approaches using pose-aware representations based on fixed body regions.

Person identification using context is another popular direction of work, where domain-specific information is exploited. Li *et al.* [74] focus on person recognition in photo-albums exploiting context at multiple levels. They propose a transductive approach where a spectral embedding of the training and test examples is used to find the nearest neighbors of a test sample. An online classifier is then trained to classify each test sample. They also exploited photo meta-data such as time-stamp and the co-occurrence of people to improve the performance. In [7, 157], meta-data and clothing information are exploited to

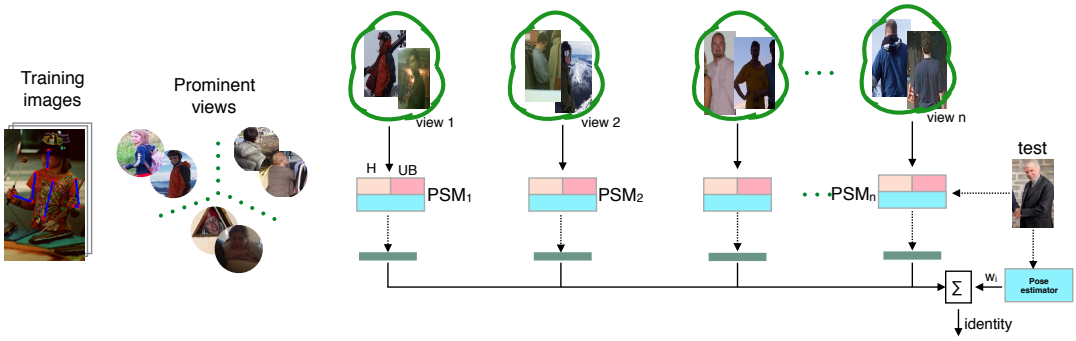


Figure 6.3 Overview of our approach: The database is partitioned into a set of prominent views (poses) based on keypoints. A *PSM* is trained for each pose based on multiple body regions. During testing, predictions from multiple classifiers, each based on a particular *PSM* representation, are obtained and combined using pose-aware weights provided by the pose estimator.

identify people in photo collections. Similarly in [39], the authors use timestamp, camera-pose, and people co-occurrence to find all the instances of a specific person from a community-contributed set of photos of a crowded public event. Sivic *et al.* [130] improve the recall by modeling the appearance of cloth, hair, skin of people in repeated shots of the same scene.

People identification in videos may use cues such as sub-title [29] or appearance models [38, 128], in addition to clothing, audio, face [136]. Similarly, a combination of jersey, face identification and contextual constraints are used to identify players in broadcast videos [13, 90].

We focus on the generic person recognition problem similar to [62, 165] that work in diverse settings without using any domain level information and demonstrate the effectiveness of the pose-aware models in different scenarios.

6.3 Pose-Aware Person Recognition

The primary challenge in person recognition is the variation in pose[1] of the subjects. The appearance of the body parts change significantly with pose. We aim to tackle this by learning pose-specific models (*PSM*s), where each *PSM* focuses on specific discriminative features that are relevant to a particular pose. We fuse the information from different *PSM*s to make an identity prediction.

Our proposed framework is shown in Figure 6.3. Given a database of training images with identity labels and key-points, we cluster the images into a set of prominent views (poses) based on keypoint

[1]We use the terms *pose* and *view* interchangeably to refer the overall orientation of the body with respect to the camera and not the location of keypoints within the body.

features. A pose estimator is learned on these clusters for view classification. For learning person representation in each view, a *PSM* is then trained for identity recognition that makes use of multiple body regions. We train multiple linear classifiers that predict the identities based on *PSM* representations. Given an input image x, we first compute the pose-specific identity scores, $s_i(y, x)$, each based on the i^{th} *PSM* representation. The final score for each identity y is a linear combination of the pose-specific scores.

$$s(y, x) = \sum_i w_i s_i(y, x), \tag{6.1}$$

where w_is are the pose-aware weights predicted by the pose estimator. To allow robustness to rare views with limited training examples, we also incorporate a `base` model in the above equation similar to [165] that is trained on the entire train set, whose scores and weights are referred as $s_o(y, x)$ and w_o respectively. The predicted label of the sample is computed as: $\arg\max_y s(y, x)$.

Our framework differs from `PIPER` in two aspects. First, our pose-aware weights are specific to each instance as opposed to the `PIPER`, which uses fixed weights computed from a validation set. Second, `PIPER` extracts features from a single model for a given localized poselet patch, however, we extract features from different pose models but combine them softly using the pose weights. This allows multiple *PSM*s that are very near in pose space (e.g.. a semi-left and left) to contribute during the prediction.

6.3.1 Learning Prominent Views

To facilitate pose-aware representations, we partition the training images into prominent views using body keypoints. Although people exhibit large variations in arm and leg positions, we consider only the informative regions such as head and torso. We construct a 24-D feature for pose clustering using 14 key points and visibility annotations as shown in Figure 6.4. It consists of -

1. 10-D *orientation feature* based on the relative location of different body parts computed as $[cos(\theta_1), \ldots, cos(\theta_8), sign(x_6 - x_3), sign(x_7 - x_4)]$, where $\theta_i, i = \{1, 2, 3, \ldots, 8\}$ denote the angle between the line joining two key points and the x-axis. For example, θ_6 is the angle between head midpoint and right shoulder. The last two elements distinguish front and back views, and are based on the sign of x-coordinate differences of left and right points of shoulder and elbow, respectively.

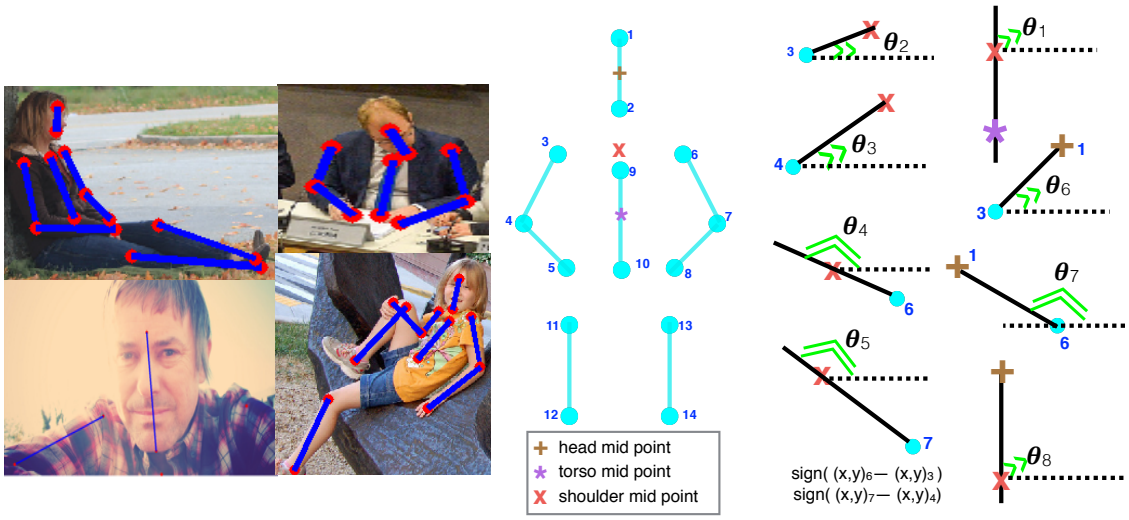


Figure 6.4 We use body keypoints (left) to learn prominent views using a set of features (right) based on orientation of informative body parts and keypoints.

2. 14-*D visibility feature*, where each element is either 1 or 0 depending upon whether the corresponding keypoint is visible or not. This provides strong pose cues as certain body parts are not visible in particular views.

To identify meaningful views, we apply k-means algorithm to cluster the images based on the above features. We first obtain a large number (30) of highly similar groups, which are then hierarchically merged to obtain seven prominent views. Few examples of our pose clusters obtained on PIPA dataset are shown in Figure 6.5. Each row from top to bottom contain images from `right`, `semi-right`, `frontal`, `semi-left`, `left`, `back` and `partial body` views. The orientation and keypoint visibility features produced tight clusters containing images with particular body orientation. The last cluster captures the instances with partial upper body such as head or shoulder, etc, in the images that are commonly seen in social media photos and movies. While we considered seven prominent views in this work, we note that generating a large number of views can be helpful, provided there are enough training samples in each cluster to train the convnets.

Once we obtain the prominent views or poses, we train an pose-estimator based on AlexNet that takes full body image as input and computes the likelihood of each pose. During testing, the pose likelihood estimated from the pose-estimator provide the pose-aware weights w_i, in Eqn. 6.1. We noticed from our experiments that, it is critical to l_2-normalize the weights to obtain the improved performance.

Figure 6.5 Pose clusters: Each row from top to bottom shows people from PIPA with particular body orientation clustered using orientation and keypoint visibility features.

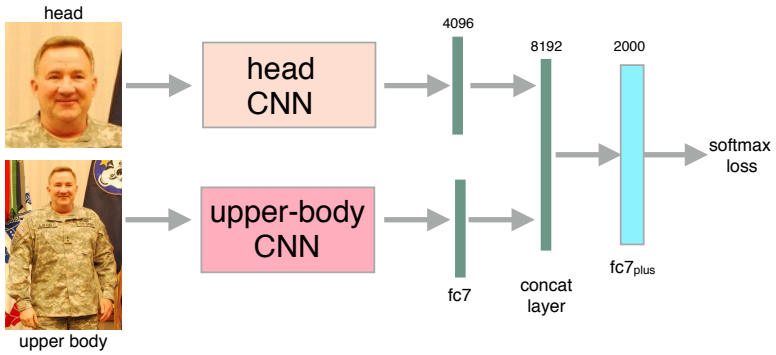


Figure 6.6 PSM: Our network consists of two AlexNets for head and upper body with a single output layer. The last fully connected layer of the two regions are concatenated and passed to an joint hidden layer with 2000 nodes.

6.3.2 Learning a PSM

To train a pose-specific model (*PSM*), we select the training samples that belong to a specific pose cluster. We consider the head and upper body regions as these are the most informative cues for recognition [62, 74]. Given a head at location (l_x, l_y) with dimensions (δ_x, δ_y), we estimate the upper body to be a box at location $(l_x - 0.5\alpha, l_y)$ of dimensions $(2\alpha, 4\alpha)$ where $\alpha = min(\delta_x, \delta_y)$.

Given different body parts, one possibility is to train independent convnets on each of these regions [62, 74, 165]. However, discriminative body regions that help in recognition may vary across training instances. For example, Figure 6.2(a)-4 contains an occluded face region and is less informative. Similarly, upper body may be less informative in some other instances. If such noisy or less informative regions influence the optimization process, it may reduce the generalization ability of the networks.

We propose an approach to improve the generalization ability by allowing the network to selectively focus on informative body regions during the training process. The idea is to optimize both the head and upper body networks jointly over a single loss function. Our *PSM* contains two AlexNets corresponding to the head and upper-body regions (see Figure 6.6). The final `fc7` layers of each region are concatenated and passed to a joint hidden layer ($\texttt{fc7}_{\text{plus}}$) with 2000 nodes before the classification layer. This provides more flexibility to the network to make the predictions based on one region even if the other region is noisy or less informative. As we show in our experiments (Table 6.5), the joint training approach performs better than separate training of regions.

Figure 6.7 Few images from (top) PIPA [165] and (bottom) Hannah [102] datasets. Faces in these datasets exhibit diverse variations in terms of occlusion, resolution, age, pose and illumination. In many instances, face is not completely visible and captured from the back-view.

6.3.3 Identity Prediction with PSMs

We derive multiple features from each *PSM* and train classifiers on these feature vectors. The primary feature vector ($\mathcal{F}$) consists of the sixth and seventh layers of the head (h) and upper body (u), and the joint fully connected layer ($\mathcal{F}$: $<\texttt{fc6}_\texttt{h}, \texttt{fc7}_\texttt{h}, \texttt{fc6}_\texttt{u}, \texttt{fc7}_\texttt{u}, \texttt{fc7}_\texttt{plus}>$). In addition to $\mathcal{F}$, we define two additional feature vectors solely based on the head and upper body layers - $\mathcal{F}_h$: $<\texttt{fc6}_\texttt{h}, \texttt{fc7}_\texttt{h}>$ and $\mathcal{F}_u$: $<\texttt{fc6}_\texttt{u}, \texttt{fc7}_\texttt{u}>$. We train linear SVM classifiers on each of the above feature vectors to obtain the identity predictions. The pose-specific identity score, $s_i(y, x)$, is simply the sum of the three SVM classifier outputs.

$$s_i(y, x) = \sum_{f \in \{\mathcal{F}, \mathcal{F}_h, \mathcal{F}_u\}} P_i(y|f; x), \tag{6.2}$$

where $P_i(y|f; x)$ is the class y score of the sample x predicted by the classifier trained on the feature f in i-th view.

6.4 Datasets and Setup

We select three datasets from the domain of photo-albums, movies and sport broadcast videos. Each of these settings have their own set of advantages and challenges as summarized in Table 6.1. To the best of our knowledge, this is the first work that evaluates person recognition in such diverse scenarios.

	PIPA [165]	Hannah [102]	Soccer
Train instances	6,443	2,385	19,813
Train subjects	581	26	28
Test instances	6,443	202,178	51,051
Test subjects	581	41	28
Annotations	Head	Face	Body
Domain variation	No	Yes	No
Clothing	Yes	No	No
Age gap	No	Yes	No
Head resolution	High	Medium	Low
Motion blur	No	Moderate	Severe
Deformation	Less	Moderate	Severe

Table 6.1 Comparison of the datasets in terms of statistics, annotations, merits and challenges.

6.4.1 Photo Album Dataset

PIPA [165] consists of 37,107 photos containing 63,188 instances of 2,356 identities collected from user-uploaded photos in Flickr. Few images from the dataset are shown in Figure 6.7. The dataset consists of four splits with an approximate ratio of 45:15:20:20. The larger split is primarily used to train convnets, second split to optimize parameters during validation and the third split to evaluate recognition algorithms. The evaluation set is further divided into two equal subsets, each with 6,443 instances belonging to 581 subjects for training and testing the classifiers. We follow PIPA experimental protocol and train the classifiers on one fold and test on another fold, and vice-versa. We also conduct experiments on challenging splits introduced by Oh *et al.* [62] based on album, time and day information.

6.4.2 Hannah Movie Dataset

We consider "Hannah and Her Sisters" dataset [102] to recognize the actors appearing in the movie. The dataset consists of 153,833 movie frames containing 245 shots and 2,002 tracks with 202,178 face bounding boxes. We regress the face annotations to get the rough estimate of head. There are a total of 254 labels of which 41 are the named subjects. The remaining labels are the unnamed characters (boy1, girl1, *etc*) and miscellaneous regions (crowd, *etc*). Few images from Hannah testset are shown in Figure 6.7.

To create the gallery set, images are collected from the IMDB profile[2] of actors appearing in the movie. Head bounding boxes are then manually annotated for each instance. Few images from IMDB

[2] `http://www.imdb.com/title/tt0091167/fullcredits`

database are shown in Figure 6.8. We often relied on text tags associated with photos for annotation whenever the photos contain multiple confusing identities. Of the 41 named characters, only 26 prominent actors had profiles in IMDB. The IMDB train set consists of $2{,}385$ images belonging to 26 prominent actors appeared in the movie. There are a total of $159{,}458$ instances belonging to these 26 actors in the test set.

There is a significant age variation between train and test instances since the Hannah instances are created from a particular year (1986) while the IMDB photos are captured over a long period of time. In addition, there is a large domain contrast between IMDB and Hannah testset in terms of lighting, camera and imaging conditions. This creates a more challenging setting to match identities between IMDB and Hannah instances.

6.4.3 Soccer Dataset

We create soccer dataset from the broadcast video of World cup 2014 final match played between Argentina and Germany. We considered only replay clips as these capture the important events of the match. We further filtered the replay clips to retain only those clips that are shot in close-up and medium views. We used VATIC toolbox [142] to annotate the players in videos. Our soccer dataset consists of 37 video clips with an average duration of 30 secs. It consists of 28 subjects with 13 players from Germany team, 14 players from Argentina team, and a referee. We show images from soccer dataset in Figure 6.9.

Unlike PIPA, we marked full-body bounding boxes for each player since head is not visible or out-of-view in many instances, and it also is difficult to estimate the bounding boxes of different body regions from head, due to large deformations. We followed PIPA annotation protocol and labeled the players regardless of their pose, resolution and visibility. We annotated the players to generate continuous tracks even in the presence of severe occlusion. Whenever it is difficult to recognize the players, we relied on additional clues such as hair, shoes, jersey number and accessories. However, we do not rely on any of these domain-specific cues in this work. For evaluation purposes, we randomly select 10 clips into training and remaining 27 clips into testing. This resulted in $19{,}813$ instances in training set and $51{,}051$ instances in testing set.

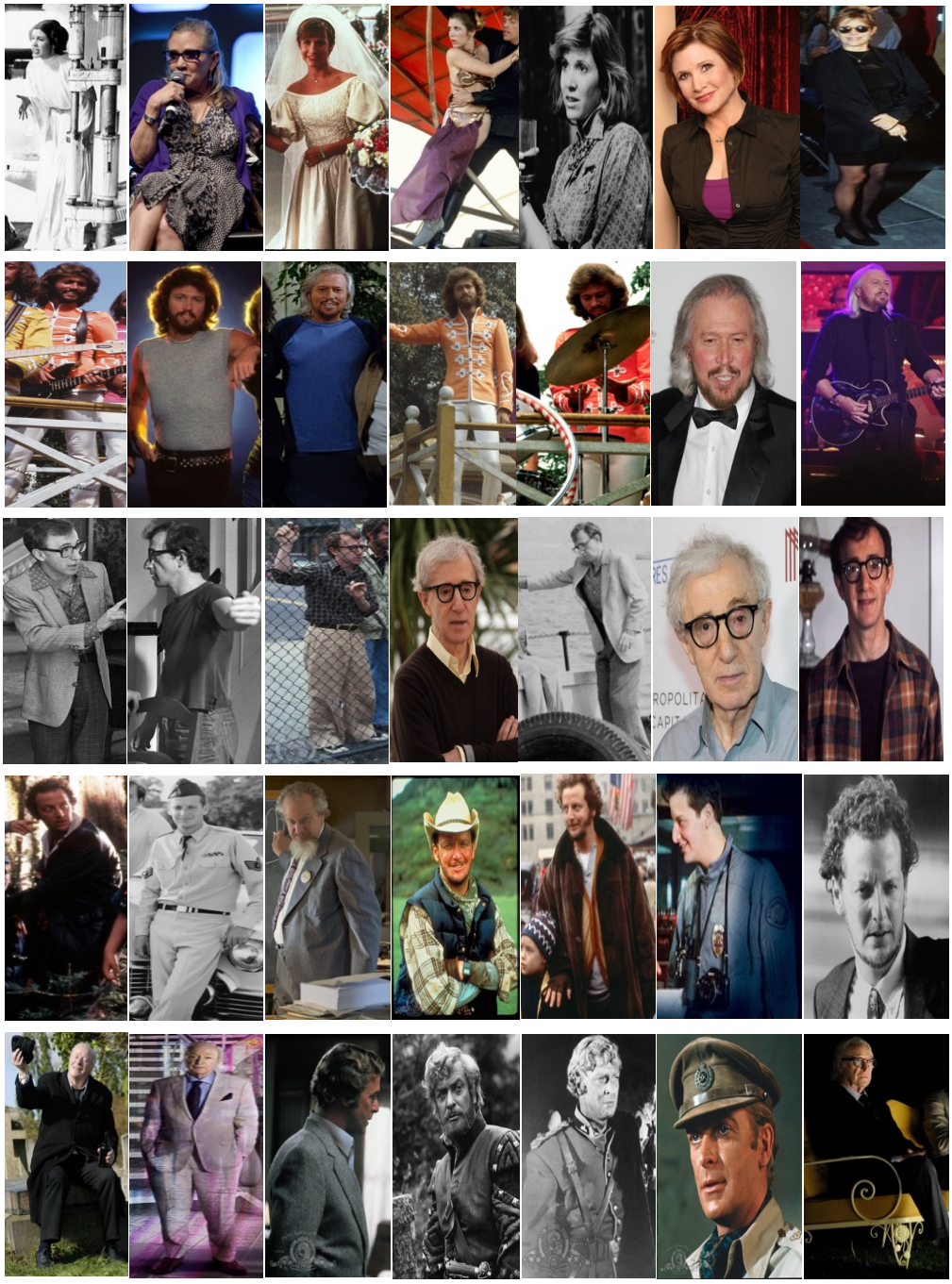


Figure 6.8 IMDB dataset. Each row shows few images of an actor from the dataset. We create IMDB gallery set to recognize actors appearing in Hannah movie. IMDB images are captured at various time periods and contain significant age variations.

Figure 6.9 Images from soccer dataset. It offers a challenging person recognition scenario due to low resolution, high occlusion, deformation and motion blur exhibited by soccer instances.

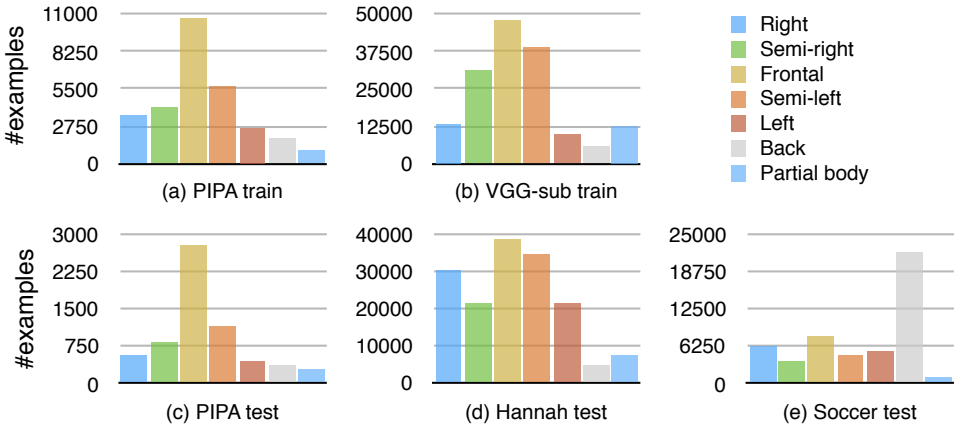


Figure 6.10 Pose statistics of different datasets.

6.5 Results and Analysis

For all our experiments, we use the *PSM* models trained on larger set of PIPA consisting of 29,223 instances. We annotate PIPA train instances with keypoint locations to learn prominent views as discussed in 6.3.1. The number of instances in each view after pose clustering is shown in Figure 6.10(a). We train a separate *PSM* on `frontal`, `semi-left`, `left`, `semi-right`, `right`, `back` and `partial` views. The `base` model is trained on the entire PIPA train set. We augment each view by horizontal flipping of the instances from its symmetrically opposite view. For instance, images in left view are flipped and augmented to right view. We use Caffe library [59] for our implementation. For optimization, we use stochastic gradient descent with a batch size of 50 and momentum coefficient of 0.9. The learning rate is initially set to 0.001, which is decreased by a factor of 10 after every 50,000 iterations. We train the networks for a total of 300,000 iterations. The parameter C is set to 1 for training SVM and the base weight w_0 to 1.

We noticed that the *PSM*s trained on view samples lead to over-fitting. To overcome this, we used a subset of VGGFace dataset [104] for initializing the networks. We extended the face annotations and selected only those instances that have full body in the VGG images. The number of examples used to initialize *PSM*s are shown in Figure 6.10(b). We make two important points regarding the additional data. First, the extra data ($\sim$160K) we considered is much smaller compared to `PIPER` ($\sim$4M faces for training DeepFace [135]) and `naeil` ($\sim$500K from four different datasets). Second, the proposed improvement is primarily due to the pose-aware combination strategy and not the ensemble of different view-specific models. This is discussed below in the Ablation study (III).

Overall performance: We compare the performance of various approaches on PIPA test splits in Table 6.2, including both baseline and contextual results of Li *et al.* [74] to provide a comprehensive review. When the contextual information is not considered, our approach outperforms all the previous approaches achieving an 89.05% on the original split. On the Hannah dataset, our approach outperforms `naeil` by a large margin as shown in Table 6.3. Finally, we show the results on the newly created player recognition in soccer in Table 6.4.

We notice on all the datasets that use of multiple body regions helps in recognition strengthening the motivation behind `PIPER` and `naeil` algorithms. We also show the results by merging track labels on Hannah and Soccer datasets to understand their impact on recognition. We reassign the frame labels based on simple majority voting of all the frames in a track. The results suggest that track information if available should be used to improve the performance.

The accuracies on Hannah and Soccer datasets are much lower compared to PIPA owing to lower resolution, motion blur, heavy occlusions, and age variations. We also observe a little improvement over `naeil` on soccer dataset due to unusual poses (kicking, falling, *etc*.) of the subjects. A large majority of these images are predicted as "back-view" by the pose-estimator (see Figure 6.10(e)). Since the performance of our `back` view model is poor due to limited training data, we noticed only minimal improvement.

Ablation study (I): We analyze the effectiveness of different features and joint optimization strategy with `base` model in Table 6.5. The use of both $\texttt{fc}_6$ and $\texttt{fc}_7$ features improve the performance for all the body regions. The head ($\mathcal{F}_h$) and upper body ($\mathcal{F}_u$) features obtained through joint training outperform the head ($\texttt{h}_2$) and upper body ($\texttt{u}_2$) features obtained through separate training by almost two percent points. Similarly, the concatenation of head and upper body features ($\mathcal{F}$) through joint training perform better than separate training ($\mathcal{S}_1$). Finally, the combination of three classifiers ($s_0(y, x)$) from head, upper-body and joint features further bring the performance improvement. We note that, our single `base` model with joint-training strategy and combination of classifiers itself outperforms `naeil`, which reports an accuracy of 86.78% with 17 models.

Ablation study (II): We conduct experiments to measure the performance of pose-specific *PSM* models using PIPA test set. In each experiment, we consider only those examples that are in i-th pose and select half of them randomly for classifier training and remaining half into testing. We extract $\mathcal{F}_u$ feature for these examples using *PSM* and `base` models. Figure 6.11 shows the performance of each model in recognizing examples from different views. It shows that for `frontal`, `semi-left`

Method	Original	Album	Time	Day
PIPER [165]	83.05	-	-	-
naeil [62]	86.78	78.72	69.29	46.61
Li *et al.* w/o context [74]	83.86	78.23	70.29	56.40
Li *et al.* with context [74]	88.75	**83.33**	**77.00**	**59.35**
Our approach	**89.05**	82.37	74.84	56.73

Table 6.2 Performance comparison (%) of various approaches on different PIPA splits.

Method	Accuracy without tracks	Accuracy with tracks
Head (H)	27.52	31.91
Face (F)	26.53	31.55
Upper body (U)	16.49	17.72
Separate training of H and U	31.86	36.10
Joint training of H and U	32.92	37.74
naeil [62]	31.41	37.57
Our approach	**40.95**	**44.46**

Table 6.3 Recognition performance (%) of various approaches on Hannah movie dataset using IMDB dictionary.

Method	Accuracy without tracks	Accuracy with tracks
Head (H)	17.68	20.54
Upper body (U)	18.01	19.76
Separate training of H and U	17.62	20.68
Joint training of H and U	18.35	20.18
naeil [62]	19.45	23.77
Our approach	**20.15**	**24.31**

Table 6.4 Performance comparison (%) on Soccer dataset.

	Feature		Accuracy
Face (F)	$\texttt{fc7}_\texttt{f}$		66.83
	$[\texttt{fc6}_\texttt{f}\ \texttt{fc7}_\texttt{f}]$		70.40
Head (H)	$\texttt{fc7}_\texttt{h}$	$\ldots(\texttt{h}_1)$	76.81
	$[\texttt{fc6}_\texttt{h}\ \texttt{fc7}_\texttt{h}]$	$\ldots(\texttt{h}_2)$	79.54
Upper body (U)	$\texttt{fc7}_\texttt{u}$	$\ldots(\texttt{u}_1)$	72.26
	$[\texttt{fc6}_\texttt{u}\ \texttt{fc7}_\texttt{u}]$	$\ldots(\texttt{u}_2)$	75.19
Separate training of H and U	$[\texttt{h}_1\ \texttt{u}_1]$		82.90
	$[\texttt{h}_2\ \texttt{u}_2]$	$\ldots(\mathcal{S}_1)$	84.01
Joint training of H and U	$\texttt{fc7}_\texttt{plus}$		85.98
	$[\texttt{fc6}_\texttt{h}\ \texttt{fc7}_\texttt{h}]$	$\ldots(\mathcal{F}_h)$	82.22
	$[\texttt{fc6}_\texttt{u}\ \texttt{fc7}_\texttt{u}]$	$\ldots(\mathcal{F}_u)$	77.62
	$[\texttt{fc7}_\texttt{h}\ \texttt{fc7}_\texttt{u}]$		85.05
	$[\texttt{j}_1\ \texttt{fc7}_\texttt{h}\ \texttt{fc7}_\texttt{u}]$		85.22
	$[\texttt{fc7}_\texttt{plus}\ \texttt{fc6}_\texttt{h}\ \texttt{fc6}_\texttt{u}]$		86.10
	$[\texttt{fc7}_\texttt{plus}\ \mathcal{F}_h\ \mathcal{F}_u]$	$\ldots(\mathcal{F})$	86.27
	$s_0(y, x)$		86.96

Table 6.5 Performance (%) of different features obtained from separate and joint training of regions on PIPA test set.

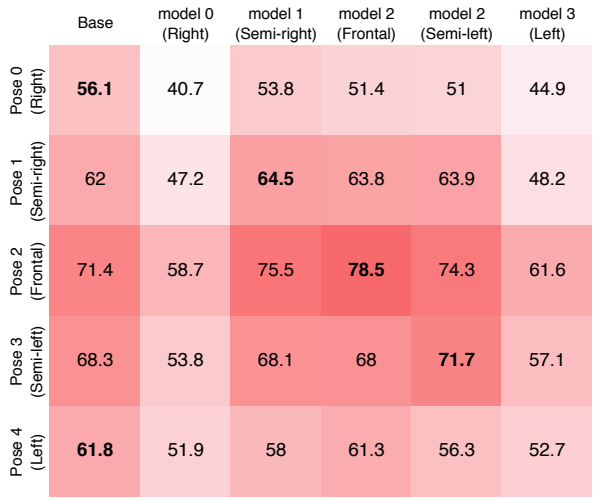


Figure 6.11 Effectiveness of *PSM*s: Each row shows the performance of test examples in a particular pose represented using the different *PSM*s. Frontal and semi-profile examples are best recognized with *PSM*s. For rare views, base model trained on large data seems to be more robust compared to corresponding *PSM*s trained on limited examples.

Fusion type	(I)	(II)
Average pooling	83.57%	87.78%
Max pooling	80.54%	85.51%
Elementwise multiplication	81.71%	86.32%
Concatenation	**84.44%**	87.62%
Pose-aware weights	–	**89.05%**

Table 6.6 Comparison of different fusion schemes for combining (i) features during joint training (using `frontal` *PSM*) and (ii) pose-aware classifier scores during testing.

and `semi-right` examples, corresponding *PSM* models outperform other models including the base model. This show that the person representations obtained from pose-specifc models are more robust than the pose-agnostic representations. However, for extreme profile views, we noticed that base model performed better than the corresponding profile models. We attribute this to the non-availability of enough profile images while training the *PSM*. For this reason we include the base model along with *PSM*s to bring more robustness when handling rear and non-prominent view images.

Ablation study (III)**:** Our approach uses two kinds of information pooling, one during *PSM* training and another for combining classifiers. For joint-training, we show the effect of different head ($\texttt{fc7}_\texttt{h}$) and upper body ($\texttt{fc7}_\texttt{u}$) combination strategies in Table 6.6 (I). The simple concatenation of $\texttt{fc7}_\texttt{h}$ and $\texttt{fc7}_\texttt{u}$ worked better, and hence considered. Similarly, we tried multiple strategies to combine the classifiers during testing. As can be seen in Table 6.6 (II), pose-aware weighting outperform other strategies including average pooling of the ensemble of classifiers.

Pose-wise recognition performance: The statistics of different poses is given in Figure 6.10 (bottom) for different datasets. Frontal images dominate PIPA due to which algorithms already achieve high performance ($>80\%$). Hannah consists of different poses in similar proportion while the soccer dataset contains majority of back view images. Consequently, we observe a low performance on these datasets. In Figure 6.12, we show pose-wise recognition performance. PIPA and Hannah have a similar trend in which frontal and semi-profile images are recognized with greater accuracies, while profile and back-views with less accuracy. The upper body seems to be less informative in case of Hannah as the clothing is completely different between Hannah and IMDB. The proportion of back views that are correctly recognized is slightly better in soccer setting due to large number of back view images in the classifier train set.

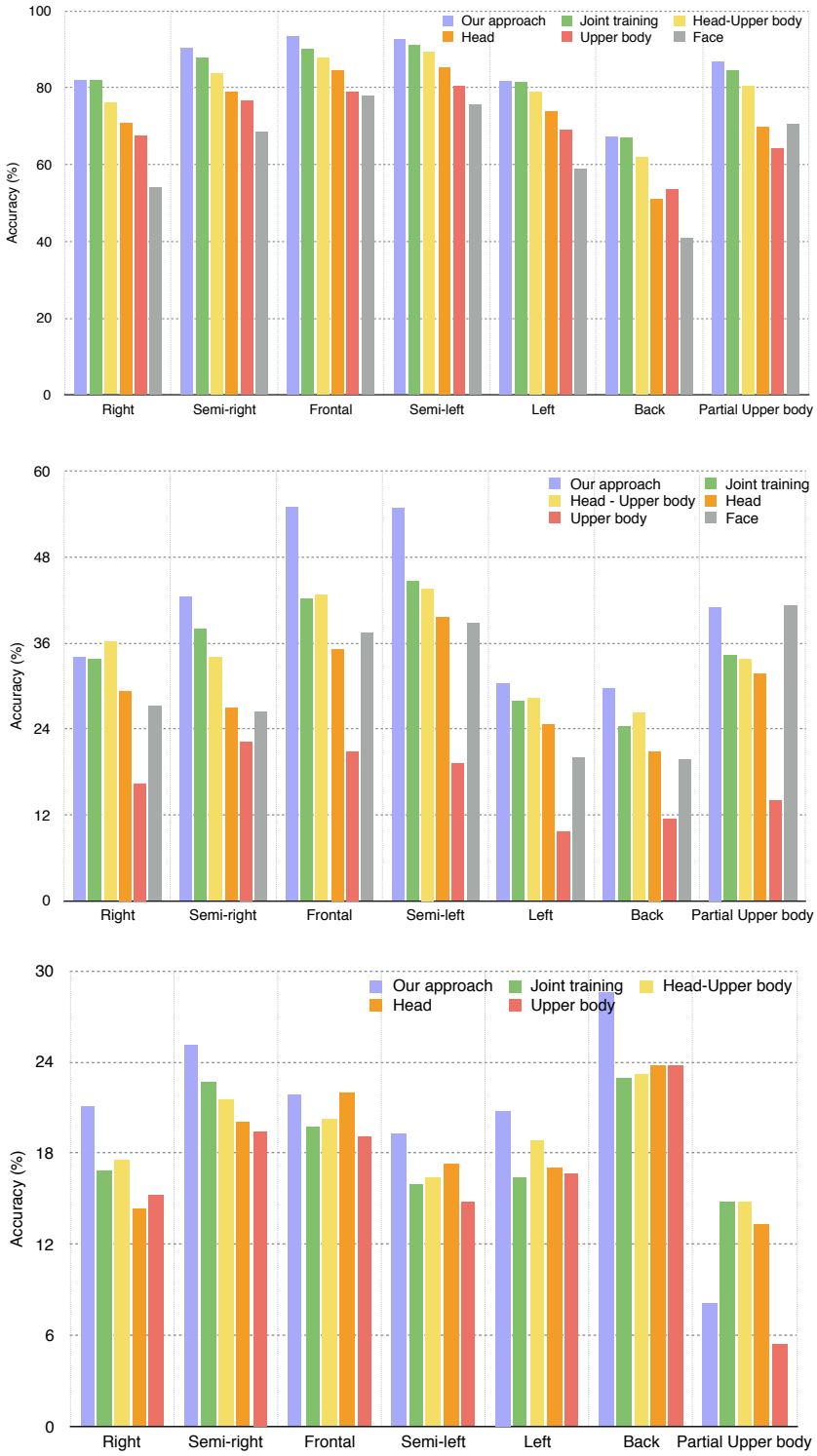


Figure 6.12 Pose-wise recognition performance on PIPA (top), Hannah (middle) and Soccer (bottom) datasets.

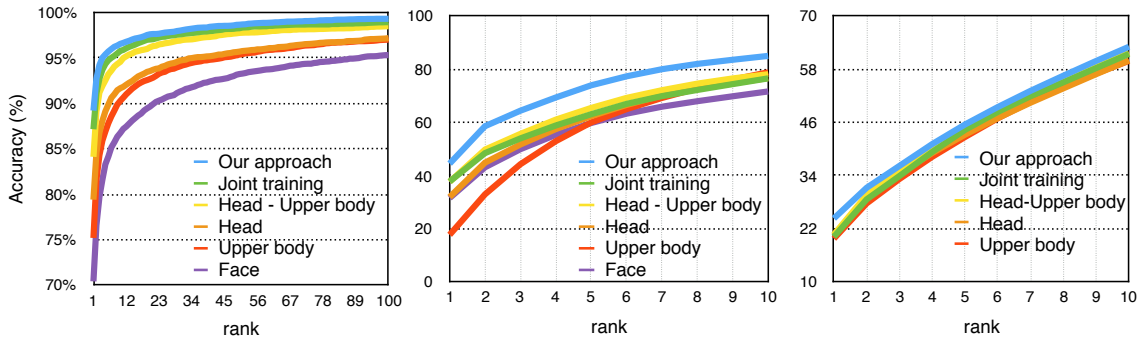


Figure 6.13 CMC curves of various approaches on PIPA (left), Hannah (middle) and Soccer (right) datasets.

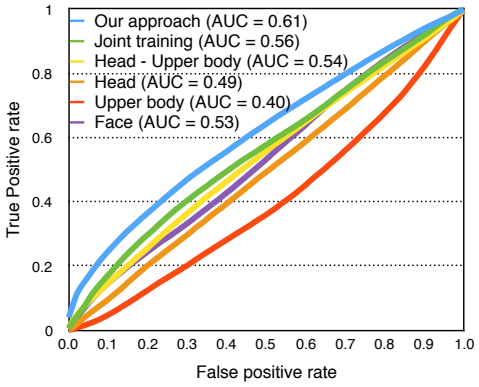


Figure 6.14 ROC curves of various approaches in rejecting unknown Hannah instances based on normalized prediction scores.

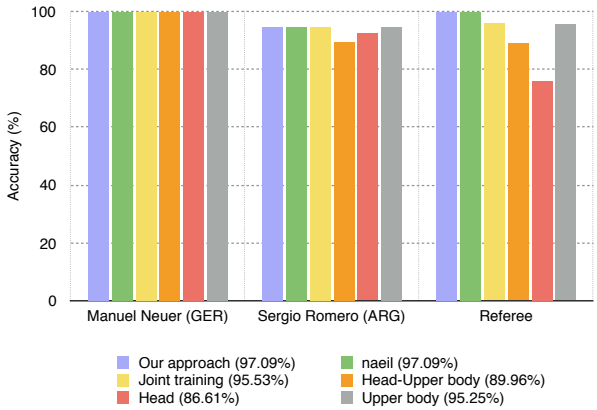


Figure 6.15 Effect of clothing on recognition.

Rank-n identification rates: The cumulative matching characteristic (CMC) curves are shown in Fig 6.13. Our approach achieves rank-10 accuracies of 96.56%, 84.83% and 63.2% on PIPA, Hannah and Soccer datasets respectively. On Hannah, we noticed a big difference of 12% between rank-1 and rank-2 performance. The performance gap between different approaches tend to reduce with higher rank.

Handling unknown instances: In the movie scenario, the test set has 41 ground truth labels while there are only 26 subjects in the trainset. Therefore, recognition algorithms should have the ability to reject such unknown instances. To achieve this, we l_2-normalized the predicted class scores and considered the maximum score as a confidence measure. The confidence score obtained on pose-aware representations are more robust in rejecting unknown, the performance being measured using ROC curve in Figure 6.14.

How informative is clothing? Though it is intuitively obvious that clothing helps in recognition, a qualitative evaluation is not done previously. We perform such a study using the soccer dataset. We show the performance of different approaches on three subjects (*Manuel Neuer*, *Sergio Romero* and *Referee*) with unique clothing in Figure 6.15. The first two subjects are the goal keepers of the Germany and Argentina, respectively.

As seen in Figure 6.15, upper body region, which is often less informative compared to head, outperforms head by a large margin due to clothing. The concatenation of head and upper body obtained through separate training is worse than upper body feature alone. On the other hand, the concatenation of features using jointly trained model is more robust and performs much better as it provide more flexibility to focus on selective regions. Finally, the overall performance of pose aware models and `naeil` are identical.

It is interesting to note that, convnets that are trained for identity recognition can distinguish clothing without any explicit modeling or hand-crafted features [90].

Recognition per subject: Figure 6.16 shows the number of images for each actor in IMDB and Hannah test sets along with their individual recognition performances. We observe that, for those subjects with sufficiently large number of training instances (*Michael Caine, Barbara Harshey, Woody Allen, Julia Louis-Dreyfus, and Mia Farrow*), the performance is high as expected. For subjects with less than 20 training instances, the performance is very low. However, whenever there is a large difference in age between train and test instances (*Carrie Fisher, Dianne West, Richard Jenkins*), the performance is poor despite having enough training examples.

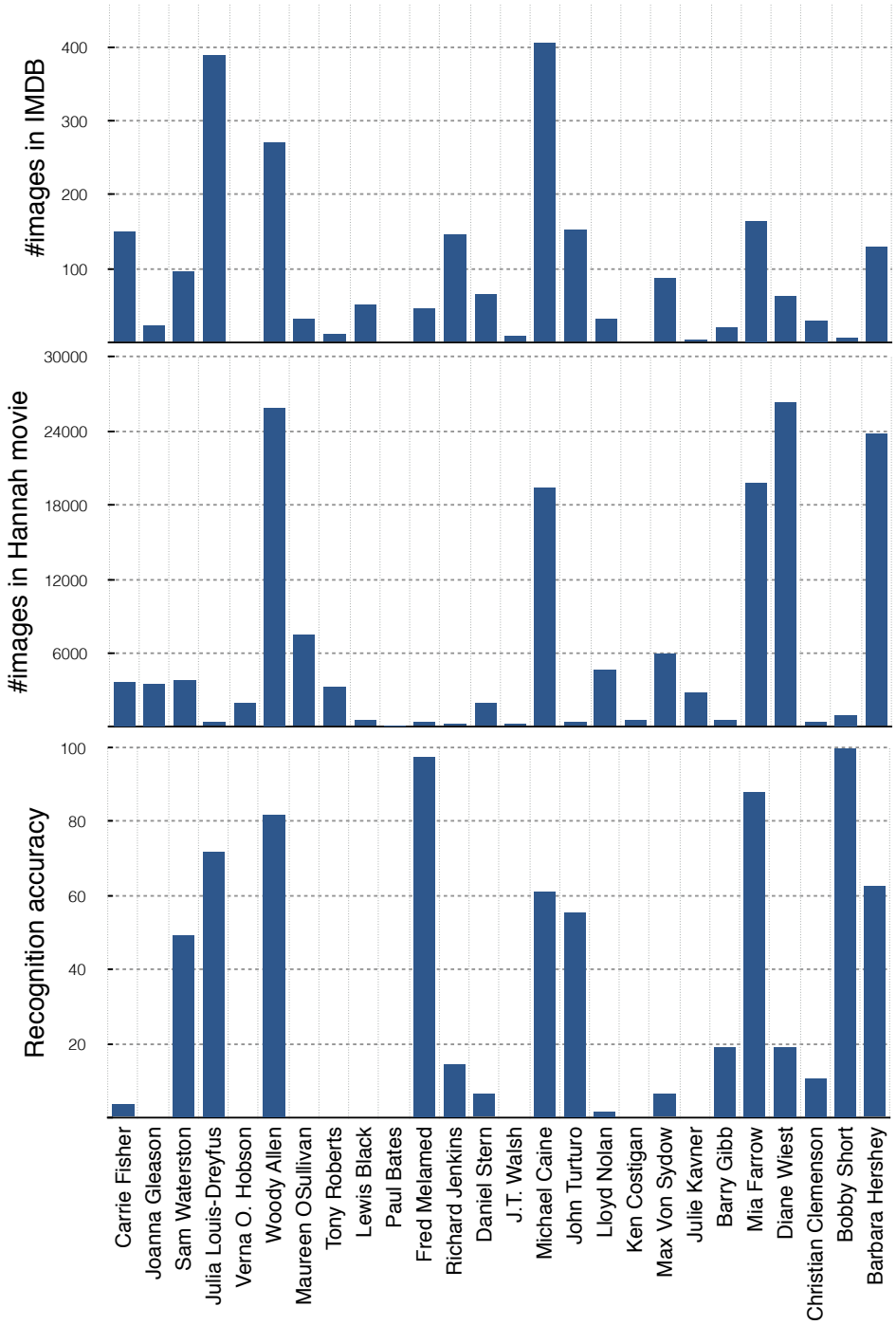


Figure 6.16 Number of images for each actor in (top) IMDB and (middle) Hannah movie test set. We show the (bottom) recognition performance of each actor on the test set.

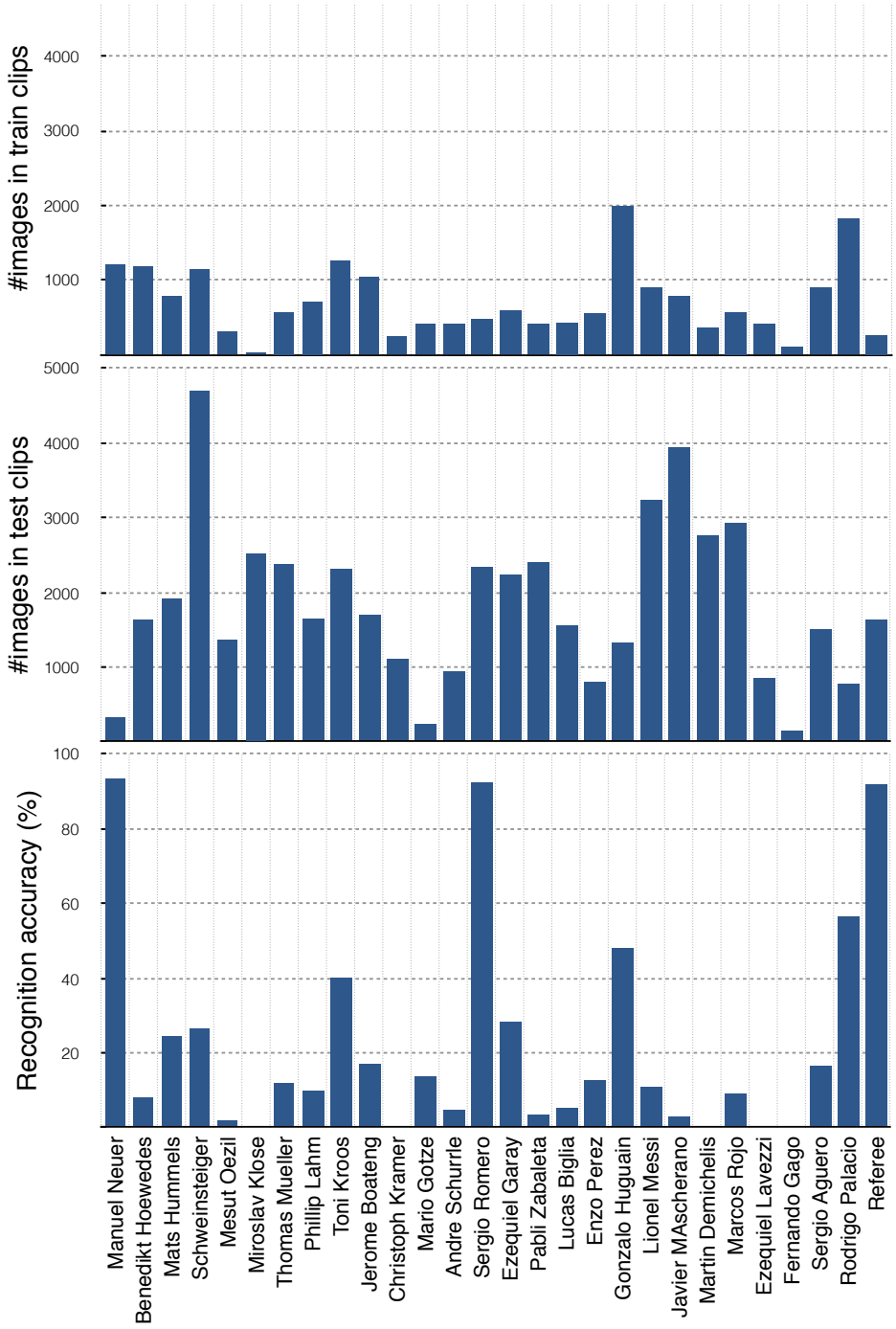


Figure 6.17 Number of images for each player in the training (top) and test (middle) split of Soccer dataset. We also show the (bottom) recognition performance of each player.

Similarly, we show the statistics of soccer players along with their individual performances in Figure 6.17. We see a similar trend of high performance for subjects (*Gonzalo Huguain* and *Rodrigo Palacio*) with sufficient training instances. We also observe a near 100% accuracy for goal keepers (*Manuel Neuer* and *Sergio Romero*) and the referee due to clothing cues, which are discussed next.

Recognition performance of top subjects: We compare the recognition performance of various approaches on 5 most occurring movie and soccer subjects in Figure 6.18 and Figure 6.19, respectively. Our approach reaches an accuracy of 61.17% on top actors, which is significantly better than `naeil`. Note that the overall performance of `naeil` with 17 models is comparable to head and upper body. Unlike photo-albums, clues such as scene and human attributes like age, glasses, and hair color are less useful in the movie setting. For actors with less change in appearance over time (*Michael Caine* and *Woody Allen*: See row three and five in Figure 6.8), face is found to be extremely informative and robust compared to head.

On the soccer dataset, the overall performance is poor for all the approaches. This suggest to develop better representations that are able to recognize people at a distance.

Confusion between identities: We show the recognition confusion matrix for Hannah and Soccer datasets in Figure 6.20 and Figure 6.21 respectively, with and without tracking. We notice two important points related to gender and clothing. As seen from Figure 6.20, female subjects are mostly getting confused with female subjects, and similarly the male subjects are confused with male subjects. In Figure 6.21, we notice that players from each team are mislabeled with the members from the same team. These studies show the effectiveness of convnets in capturing human attributes without any explicit training. Finally, majority voting over a track helps to produce consistent predictions.

Computational complexity: The number of features extracted from each *PSM* is $18{,}384$ ($4096 \times 4 + 2000$). With 7 pose-aware models and a base model, our total feature dimension is $(18,384 \times 8)$ which is $\sim$3 times smaller than `PIPER` (4096×109) and $\sim$2 times larger than `naeil` (4096×17). For memory critical applications, $\texttt{fc7}_{\text{plus}}$ alone can be used as feature. We achieve an accuracy of 87.01% on PIPA with $\texttt{fc7}_{\text{plus}}$ still outperforming `naeil` with a feature dimension of just $16{,}000$ (2000×8).

Qualitative results: We show some qualitative results in Figures 6.22 to 6.26. Figure 6.22 shows the success and failure cases of joint training and separate training of body regions. We notice an over-influence of clothing while using separately trained and concatenated regional features, compared to the jointly training features. In Figure 6.23, we show the effectiveness of using multiple classifiers from each *PSM*. As seen in the figure, the concatenated head and upper body features ($\mathcal{F}$) may predict

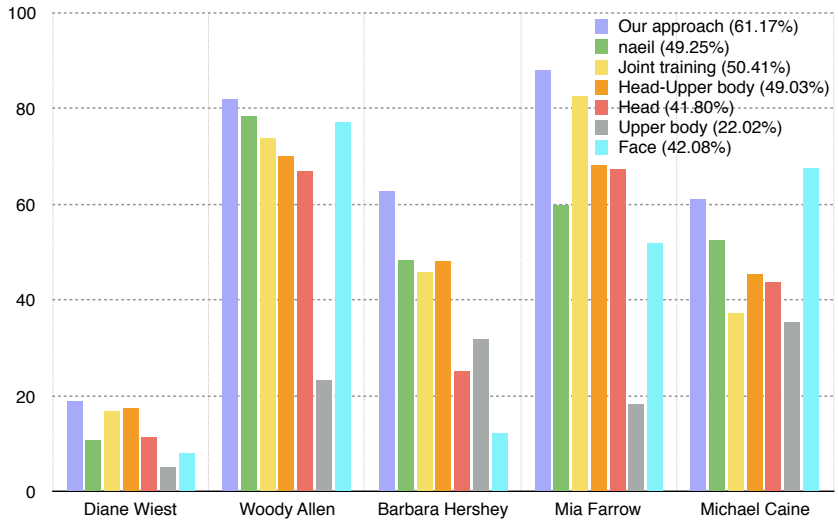


Figure 6.18 Recognition performance of five lead actors in Hannah dataset.

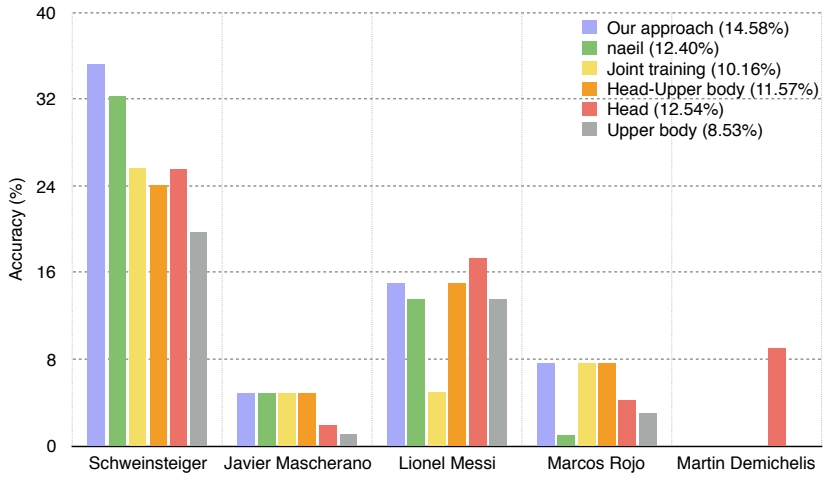


Figure 6.19 Recognition performance of five most occurring players in Soccer dataset.

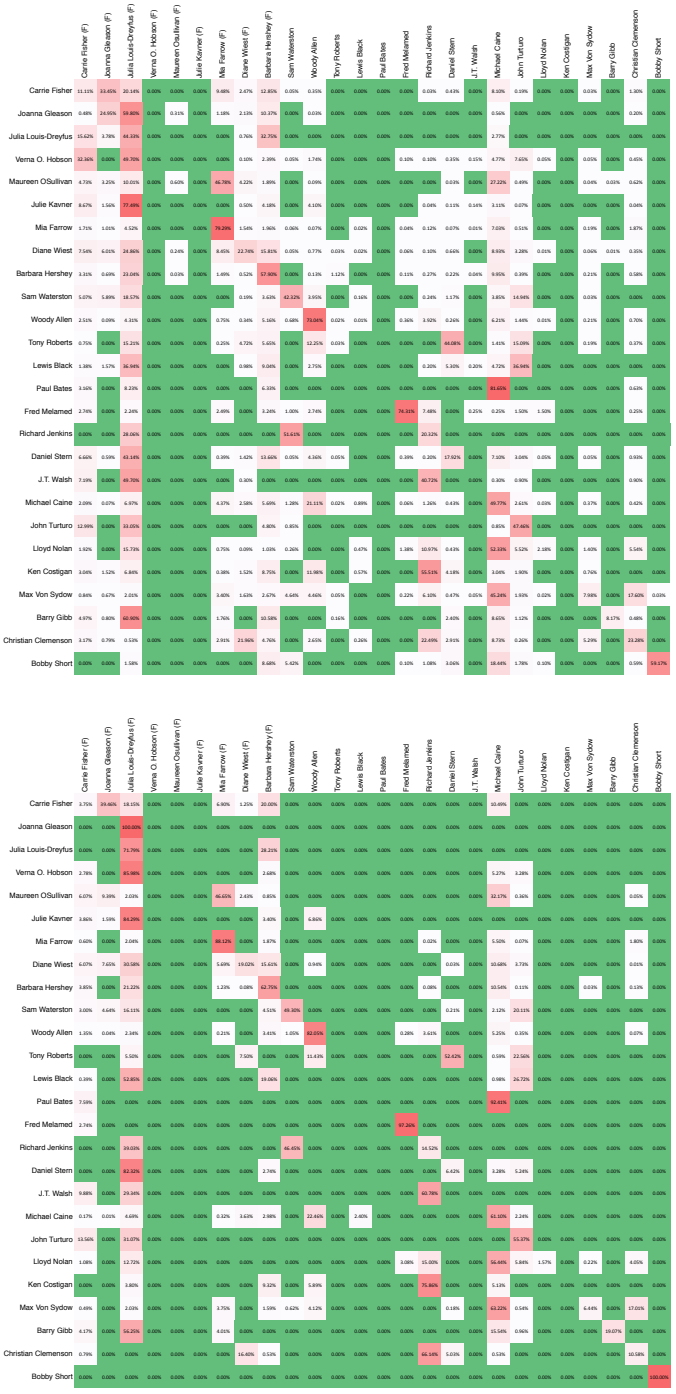


Figure 6.20 Confusion matrix on Hannah dataset (top) with and (bottom) without tracking. Best viewed on a computer after zooming.

Figure 6.21 Confusion matrix on Soccer dataset (top) with and (bottom) without tracking. Best viewed on a computer after zooming.

Figure 6.22 Success and failure cases of separate and joint training of body regions on PIPA dataset. Column one shows the test images and the column two shows the training images belonging to the predicted subject. (Left) shows the success and failure case of joint training (JT) and separate training (ST), respectively and the reverse is shown in (right).

incorrect labels even when one (or two) of these features predict correctly, due to the over influence of less informative body region. Combining these three features is found to be more robust.

We show the top scoring predictions obtained from each pose-specific *PSM* in Figure 6.24. It clearly shows how each *PSM* helps in the prediction of instances in that particular pose when the base model is unable to predict correctly. Finally, we show the success and failure cases of our approach on Hannah and Soccer datasets in Figure 6.25 and Figure 6.26 respectively, and compare with the `naeil`.

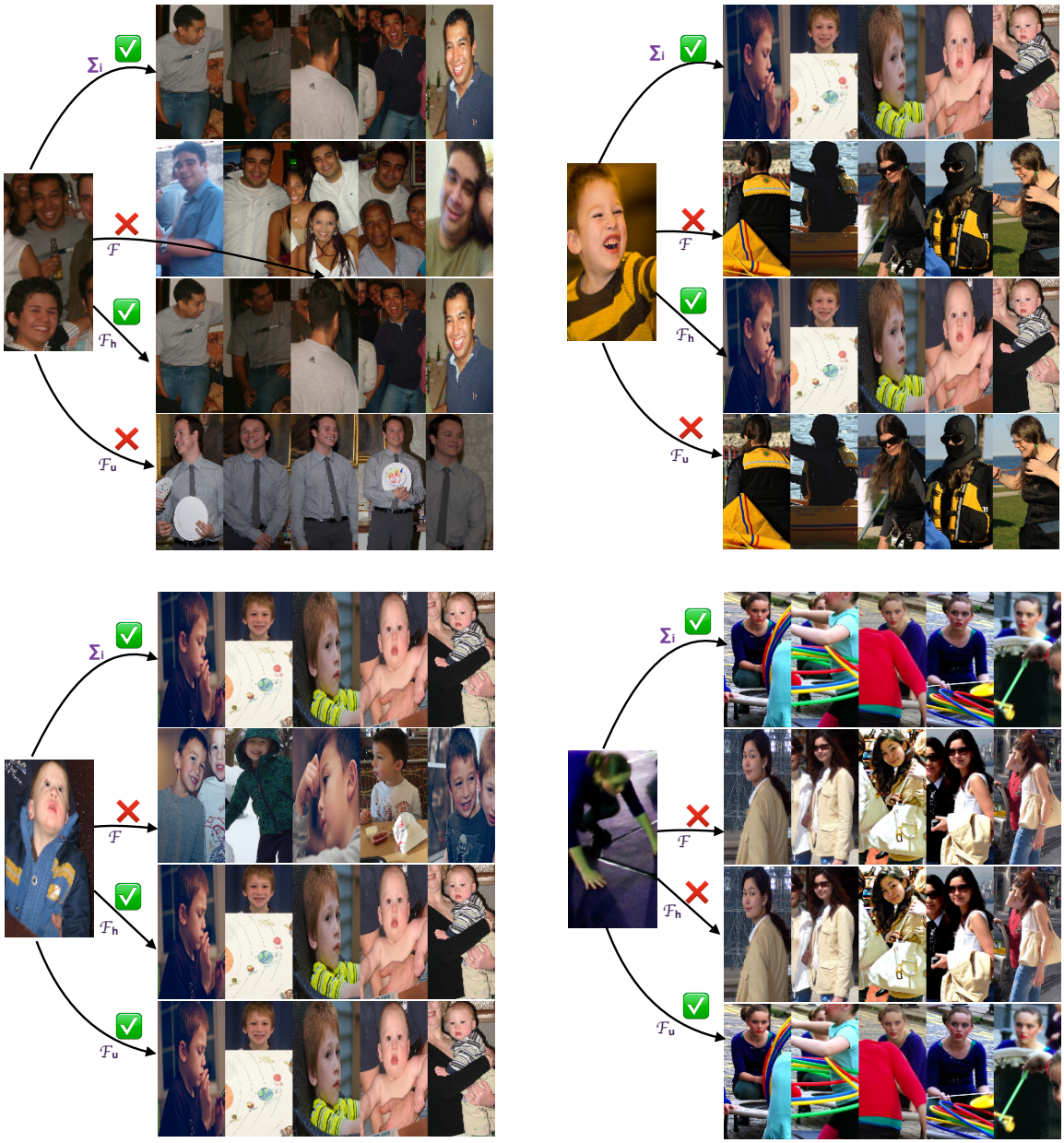


Figure 6.23 Effectiveness of multiple classifiers from each *PSM*: Column one shows the PIPA test images and the column two shows the training images belonging to the predicted subject using different approaches. The four approaches considered are the classifiers trained on head ($\mathcal{F}_h$) and upper body ($\mathcal{F}_u$) features, a classifier trained on concatenated head and upper body ($\mathcal{F}$) feature, and linear combination of three classifiers ($\sum_i$) trained on these features. It clearly shows that it is advantageous to consider individual classifiers trained on regional features and their combination for improved performance.

Figure 6.24 Success cases of pose-specific models (*PSM*s) on PIPA dataset. Each row shows the success predictions of our approach where the improvement is obtained primarily due to the specific-pose model i.e., `base` model wrongly predicts but `base` + correct `PSM` predicts correctly. Green and yellow boxes indicate the success and failure result of `naeil` respectively.

Figure 6.25 Comparison of our approach with `naeil` on Hannah dataset. Column one shows the test images and the column two shows the training images belonging to the predicted subject. (Left) in green shows the success case of our approach and the failure case of `naeil`. (Right) in red shows the failure case of our approach and the success case of `naeil`.

Figure 6.26 Comparison of our approach with `naeil` on Soccer dataset. Column one shows the test images and the column two shows the training images belonging to the predicted subject. (Left) in green shows the success case of our approach and the failure case of `naeil`. (Right) in red shows the failure case of our approach and the success case of `naeil`.

6.6 Summary

In this chapter, we focus on person recognition with multiple body cues. We show that learning a pose-specific person representation helps to better capture the discriminative features in different poses. A pose-aware fusion strategy is proposed to combine the classifiers using weights obtained from a pose estimator. The person representations obtained using a joint optimization strategy is shown to be more powerful compared to separate training of body regions. We show state-of-the-art results on three different datasets from photo-albums, movies and sport domains.

Chapter 7

End-To-End Person Recognition

7.1 Introduction

In the previous chapter, we showed the importance of multiple body cues for recognition in challenging conditions whenever face is not completely reliable. We also measured the performance of several other person recognition systems [165, 62, 74] that are based on multi-body feature extraction. In this chapter, we address a major practical bottleneck associated with these approaches.

All these multi-region recognition schemes including our approach proposed in the previous chapter follow a multi-step process, which consists of feature extraction from multiple body regions followed by their aggregation at the feature or decision stage. Several body regions are regressed from head ground truth of an image and a region convolutional neural network (convnet) is trained for each of these regions as shown in Figure 7.1(a). During testing, features are extracted from region convnets and the results are combined.

There are two shortcomings with this approach. The first is related to the training of an ensemble of models. It is not only sub-optimal to train several region models that are aimed at the common task of identity prediction but is also slower and require a large number of parameters. For instance, `PIPER` [165] and `naeil` [62] train 107 and 17 models with $\sim$6 billion and $\sim$1 billion parameters, respectively, corresponding to different body patches and attributes. The use of multiple models may also make the solution unsuitable for memory-constrained applications. The second issue is related to information fusion from multiple features or prediction scores. A stand-alone fusion scheme based on naive pooling such as concatenation, average pooling or weights estimated from validation sets may not be effective as it weighs each body feature constantly irrespective of whether it is informative or not. In an ideal case, weights should be adaptive and based on the discriminative ability of each patch.

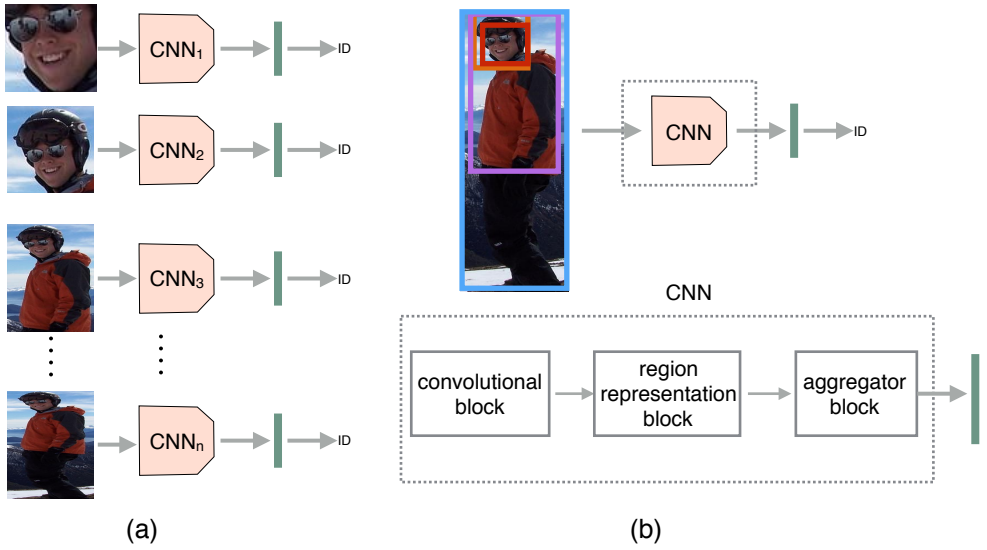


Figure 7.1 (a) Current approaches train an ensemble of models where each model focuses on a specific body region. The resulting features or classifier scores are combined during testing to determine the identity (b) We train a single end-to-end model that shares computations across multiple body regions. It produces a single discriminative person representation by adaptively aggregating pooled features.

To overcome these issues, we propose an end-to-end approach for person recognition in which a single convolutional neural network is designed to produce a compact person representation. Our architecture consists of three essential blocks as shown in Figure 7.1(b). The first block takes a complete person image as input and produces convolutional feature maps. The convolutional maps are *shared* across multiple body regions and hence require only a single forward pass to compute. The second block pools features from multiple locations on the convolutional maps based on input region of interests (ROI) and generate image representation for each body region. Finally, the third block determines the weights for these region representations using an attention mechanism with a parameter that can be trained along with other network parameters. The representations of individual regions are then combined to produce a compact and discriminative person representation.

Our work draws inspiration from region pooling in object detection (R-CNN family [42, 51]) and attention mechanism in memory networks [131]. We apply these techniques originally proposed for completely different problems to achieve end-to-end person recognition. While region pooling selects features from multiple object hypotheses in R-CNN [42], we use it to pool features from several overlapping regions of the same individual. Our work shows another effective application of region pooling in fine-grained classification problem. Similarly attention mechanisms that were successful for model-

ing long-term dependencies in sequential tasks is used to combine multiple region representations into a single person representation.

We evaluate our proposed approach on three challenging person recognition datasets and show several advantages of our proposed approach. A single model produces recognition performance that is significantly better than the baseline methods relying on large ensemble of models and comparable to the state-of-the-art approaches. Our approach requires far less parameters and is computationally faster due to sharing of features across regions.

7.2 Related Work

Person recognition that uses face together with additional cues is getting increased attention in the recent years. In addition to face, domain-specific cues such as meta-data, clothing, skin and hair [7, 157, 130], sub-titles [29], audio [136], camera pose and timestamps [39] and sport jersey numbers [13, 90] are exploited to improve the performance. Though the problem has been existing for a while, it has received renewed interest in the community with the introduction of large scale person recognition dataset named PIPA [165]. Since its introduction, several approaches [62, 70, 74, 81, 88, 165] have been proposed for person recognition with additional cues.

While `PIPER` [165] trains as many as 107 models with poselet patches, `naeil` [62] uses features from 17 models trained for identity and attribute predictions. Similarly, Vijay *et al.* [70] train several pose-specific models to learn pose specific representations. An open-set compatible loss function is proposed in [88] instead of traditional softmax loss to optimizes cosine distance between samples. Contextual information existing between instances in family albums are exploited in [74, 81] to improve the performance. Compared to these approaches, ours is a more practical solution that does not require additional information or use multiple models and is computationally faster.

Person re-identification deals with matching pedestrians captured from non-overlapping camera views in video-surveillance applications. Existing works primarily focus on metric learning [44, 53] with hand-crafted [31, 86, 91] or deep learning [5, 80, 170] based features to handle variations related to view-point, pose and appearance.

Deep learning has achieved tremendous success in the recent years progressing immensely the "three R's of computer vision" namely: recognition, reconstruction and reorganization [92]. A wide variety of techniques have been proposed to solve these problems. Our work draws inspiration from

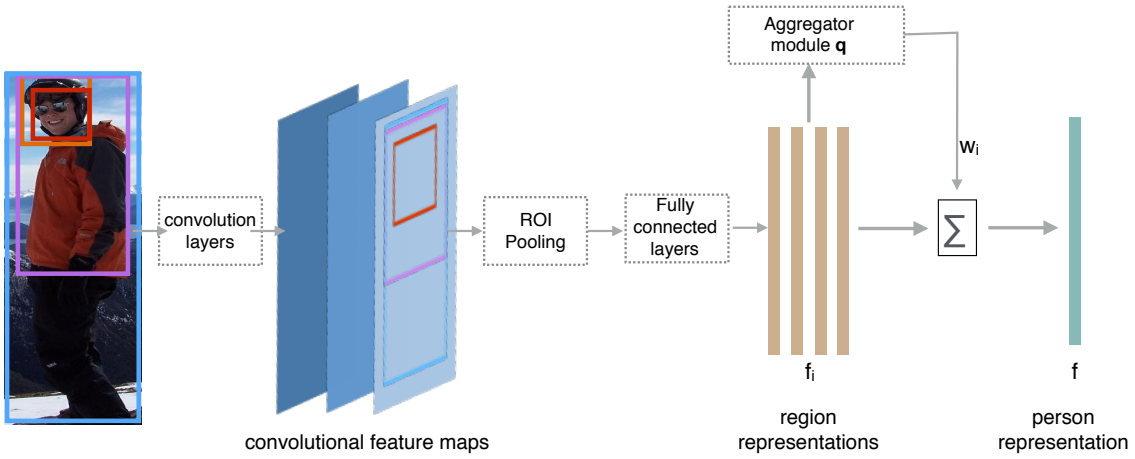


Figure 7.2 Our proposed end-to-end person recognition architecture. The input person image along with locations of body regions are passed through a series of convolution-relu layers to generates shared feature maps. Multiple features are pooled from various body locations to obtain representations for each region. The region representations are then aggregated based on adaptive weights obtained using attention mechanism by the aggregator module to produce a compact person representation.

region pooling (SPPNet [51], R-CNN [42, 112]) and attention mechanism schemes [131] used in object detection and sequential tasks, respectively. Our work demonstrates that region pooling scheme is equally effective for fine-grained classification tasks, and can be applied to extract discriminative features from different object parts. The attention schemes in [131] weighs the external memory inputs with respect to a query in sequential tasks. However, in our case the inputs do not have any order and the query is a learnable parameter that estimates the weights for different regions.

Very recently, such attention schemes are applied to aggregate facial features in the context of face recognition in videos [154] and image sets [89]. In these approaches, a scoring module generates the weights that corresponds to quality of each face in a video or image set. We finally note that our approach has some resemblance to bilinear pooling [85] in fine-grained visual recognition where features from two regions of an object are pooled to produce a bilinear vector representation.

7.3 Overview

Person recognition problem involves two essential ingredients. The first involves feature representation learning for individual body regions while the latter deals with feature aggregation such that final representation is more discriminative and produces improved performance over individual representa-

tions. Most of the existing works treat these separately as a two stage process, and focus mostly on feature representation adopting a naive pooling strategy for aggregation. To perform feature extraction for different body regions, a separate convolutional network (convnet) is trained for each body region resulting in large ensemble of models.

In this work, we consider representation generation of different body regions and their aggregation as a single process and propose an end-to-end approach to obtain a final representation. Our approach learns feature representation for multiple body regions simultaneously, and also aggregates them in an end-to-end fashion to produce a compact and discriminative person representation. The approach is simple yet powerful and requires far less parameters compared to competing models. It also avoids training of an ensemble of models corresponding to different body regions.

7.3.1 End-To-End Architecture

Our architecture is shown in Figure 7.2. The network architecture is mostly based on AlexNet [66]. While more deeper or recent architectures are equally applicable for the task, we have chosen the same architecture as previous approaches [62, 70, 165] to make the comparisons more meaningful. We however make certain modifications to perform end-to-end person recognition.

The input to our network consists of an image containing full person image, an identity label and the bounding box locations of pre-determined body regions. The network consists of three components namely: a *convolution* block that produces convolution features maps for the input image, a *region representation* block that generates representations for each region, and an *aggregator* module that generates linear weights to aggregate the regional representations using attention mechanism to produce a person representation.

7.3.2 Body Regions

An important consideration in person recognition is the choice of body regions. Current works have considered body regions either based on poselet patches [165], roughly estimated from head bounding box [62, 70, 74] or predicted using trained body part detectors [88]. Similar to `naeil` [62], we estimate body regions based on head ground-truths. Given head co-ordinates (x_h, y_h, w_h, h_h), we obtain the co-ordinates corresponding to upper-body and body regions as $(x_h - 0.5l, y_h, 2l, 3l)$ and $(x_h - 0.5l, y_h, 2l, 4l)$, respectively with $l = min(w_h, h_h)$. Face region is obtained through a face detector or obtained through regression from head ground-truths when not detected.

7.3.3 Region Representations

Based on the observation that image patches that are considered for prediction contain common regions (i.e. upper body also contains head and facial regions), it is possible to share computations across regions. Inspired from the success of region pooling in Fast-RCNN object detectors [42], we propose to train a single model for feature extraction from different body regions, thereby reducing the number of parameters required significantly. This also results in reduced training and testing times as it avoids forward and backward pass for every image patch. Our experiments demonstrate that region pooling is effective and can achieve performance on par with ensemble of individual region models.

We follow the same design principle as Fast R-CNN [42] and replace the final max pooling layer `pool3` with the ROI pooling layer. Given an input image of dimension $W \times H$, the convolution block with five convolution-relu and two max pool operations produces feature map of size $(W/17, H/17)$. These feature maps are common and shared across different body regions. The ROI pooling layer then takes the features inside the bounding box of different body regions and converts them to fixed sized feature maps by applying adaptive max pooling [42]. The pooling window size is adjusted according to the ratio of ROI region and the desired output size. We choose the output size to be 6×6. The pooled features are then passed through a region representation block consisting of three fully connected layers with an output dimensions of 4096, 4096 and 2000 respectively to obtain final representation for each region.

7.3.4 Adaptive Feature Aggregation

Once the representations are computed for different body regions, the next task is to combine them using a pooling strategy. There are two approaches that are commonly employed for aggregation. One is to combine features using naive pooling strategies such as max pooling, concatenation or average pooling and then train an identity classifier [62, 70, 74] on the aggregated feature vector. The second option is to combine the prediction scores of identity classifiers that are trained for individual body regions. The weights used for combination in such case are fixed and pre-determined using a validation set [88, 165].

However, both these approaches do not consider the following aspect of person instances in real-world images. They contain a high degree of variability in pose, view-point and occlusion and the discriminative ability of various body parts in these images vary considerably across instances. For

Figure 7.3 Examples showing adaptive weights of instances obtained during training on PIPA. Notice how weights corresponds to discriminative ability of the image patches.

example, head is more useful when upper body is occluded or invisible, while the upper body may be more useful in back-view images. An ideal approach to combine the features is to follow an adaptive scheme that gives more weightage to informative regions and less weightage to non-informative regions in a given person image.

Based on this motivation, we propose an adaptive weighting scheme to linearly combine different part level features from person instances to produce a compact and discriminative representation. We achieve this through an *aggregator* neural block used in conjunction with the convolution and region representation blocks. The parameters of the aggregator block are trainable and learned along with other network parameters in an end-to-end training process.

Formally, we are given a set of n features $S = \{f_1, f_2, \ldots, f_n\}$ extracted from n body regions of a person image and the task is to obtain a set of linear weights w_i such that aggregate feature is obtained as

$$f = \sum_{i=1}^{n} w_i f_i \tag{7.1}$$

The main objective of the aggregator module then is to generate the weights w_i using a trainable parameter vector, q. Similar to attention mechanisms followed in memory networks [131], a dot product is performed between q and each normalized feature vector $\hat{f}_i$ to obtain a set of activations a_i. A softmax operation is then applied on these activations to convert them to positive values that sum to one. The

final feature vector, w_i, is computed as:

$$a_i = q^T \hat{f}_i, \text{ where } \hat{f}_i = \frac{f_i}{||f_i||_2} \tag{7.2}$$

$$w_i = \frac{e^{a_i}}{\sum_j e^{a_j}} \tag{7.3}$$

During training, the gradients of the loss function L with respect to the parameter vector q can be obtained as $\sum_i \frac{\partial L}{\partial a_i} \frac{1}{||f_i||_2}$.

The weights w_is of some of the images obtained during training are shown in Figure 7.3. It is clear from the examples that the weights correspond to the discriminative ability of the body regions in an image. Whenever a particular region is occluded or not clearly visible, the corresponding weights are small. For images that have visible head region with near frontal pose, the head and face weights are larger compared to other regions. Also, upper body and body regions tend to get large values only when the head is less informative (e.g., $2nd$, $4th$ and $8th$ images in the first row). As expected, the weights of face and head region are highly correlated and so are the upper body and body weights as they contain large overlapping regions. Finally, we visualize several images with very large and small weights for each region in Figure 7.4.

7.4 Experiments

In this section, we conduct experiments on three person recognition benchmarks of PIPA, Hannah movie and soccer dataset and report the results. The details of the datasets and experimental setup are same as provided in the previous chapter.

7.4.1 Implementation

As mentioned earlier, our architecture is based on AlexNet [66] and trained on PIPA train set consisting of 29,223 instances. Since these datasets provide head ground-truths, we roughly estimate full body crops similar to [62]. Given head co-ordinates (x_h, y_h, w_h, h_h), we obtain co-ordinates of person crop as $(x_h - 0.5l, y_h, 2l, 5l)$, where $l = min(w_h, h_h)$. We resize the person images to 900×450 and ensure that head has a dimension of 224×224. We perform random horizontal flipping for data augmentation.

Our implementation is based on Caffe library [59]. We optimize for softmax loss using stochastic gradient descent with a batch size of 25 and momentum coefficient of 0.9. The learning rate is initially

Figure 7.4 Example of images from PIPA dataset with largest and smallest weights for each of the four body regions. Our approach clearly gives a high weightage to a body region when it is completely visible and of better quality, and a low weightage when that part is not visible or partially occluded. (best viewed by zooming on a computer)

Method	Feature	Accuracy (%)
Head	h	76.81
Face	f	66.83
Upper body	u	72.26
Body	b	69.63
Person	E = h-f-u-b	85.33
PIPER [165]	$\mathcal{P}$	83.05
naeil [62]	$\mathcal{N}$	86.78
RCNN	$\mathtt{h}_p$	75.38
	$\mathtt{f}_p$	67.57
	$\mathtt{u}_p$	72.45
	$\mathtt{b}_p$	71.20
RCNN+Avg	$\mathtt{h}_p$-$\mathtt{u}_p$	84.46
	$\mathtt{f}_p$-$\mathtt{h}_p$-$\mathtt{u}_p$	84.84
	$\mathtt{h}_p$-$\mathtt{f}_p$-$\mathtt{u}_p$-$\mathtt{b}_p$	85.51

Table 7.1 Recognition performance comparison of several separately trained features and their naive combination with region pooling on PIPA test set. $\mathcal{P}$ and $\mathcal{N}$ refer to complete PIPER and naeil feature set and subscript p refer to pooled feature.

set to 0.001 and decreased by a factor of 10 after every $50{,}000$ steps. We train the network for a total of $200{,}000$ iterations. During testing, we train an SVM classifier for prediction with parameter C set to 1.

Single vs multiple models: In Table 7.1, we first show the results of various person features trained separately using AlexNet model. To highlight the effectiveness of region pooling, we do not aggregate the features during training and compute the loss for every body region. The features are simply averaged during testing. We refer this setup as RCNN+Avg in the table. We also report the results for different combinations of regions

Table 7.1 shows that an improvement of almost 10 percent points (pp) is possible when features from head h, face f, upper body u and body b are combined as observed in [62, 165]. The same body regions when trained with pooling strategy produce similar or better performance. However, a single model trained on multiple-body regions achieves 85.51% which is comparable to the performance (85.33%) achieved using ensemble of models (E).

7.4.2 Results on PIPA Dataset

Naive vs adaptive pooling: We next experiment with various element-wise pooling strategies and compare with our adaptive weights obtained using attention mechanism in Table 7.2. As expected, fea-

Method	Feature	Accuracy (%)
RCNN+Max	$\mathtt{h}_p$-$\mathtt{f}_p$-$\mathtt{u}_p$-$\mathtt{b}_p$	83.56
RCNN+Mul	$\mathtt{h}_p$-$\mathtt{f}_p$-$\mathtt{u}_p$-$\mathtt{b}_p$	84.01
RCNN+Avg	$\mathtt{h}_p$-$\mathtt{f}_p$-$\mathtt{u}_p$-$\mathtt{b}_p$	85.51
RCNN+Conc	$\mathtt{h}_p$-$\mathtt{f}_p$-$\mathtt{u}_p$-$\mathtt{b}_p$	86.15
RCNN+Adap (N2NPR)	$\mathtt{h}_p$-$\mathtt{f}_p$-$\mathtt{u}_p$-$\mathtt{b}_p$	**87.23**

Table 7.2 Performance comparison of our proposed end-to-end recognition with various pooling strategies.

Method	Feature	Original	Album	Time	Day
Head	h	76.81	67.48	57.03	36.31
Face	f	66.83	66.83	59.87	63.78
Upper body	u	72.26			
Body	b	69.63	59.20	44.89	20.45
Person	E	85.33	76.52	66.34	42.10
PIPER [165]	$\mathcal{P}$	83.05	-	-	-
naeil [62]	$\mathcal{N}$	86.78	78.72	69.29	46.61
N2NPR	$\mathtt{R}^{\dagger}{}_p$	**87.23**	**80.98**	**71.52**	**50.23**

Table 7.3 Performance comparison (%) of various features, and their combination with our proposed approach N2NPR on different PIPA test splits. $^{\dagger}\mathtt{R}_p$ = $\mathtt{h}_p$-$\mathtt{f}_p$-$\mathtt{u}_p$-$\mathtt{b}_p$.

ture aggregation by max pooling (RCNN+Max) produces the least performance since it throws away information. While concatenation (RCNN+Conc) produces better performance than average pooling (RCNN+Avg), it is not feasible since it increases number of parameters with body regions. And finally adaptive pooling (RCNN+Adap) using our proposed attention mechanism improves over average pooling and performs better than all other strategies.

Performance of end-to-end model: Table 7.3 shows the results of N2NPR on different PIPA test splits. Our approach obtains significant improvement over naeil which relies on ensemble of region models.

Comparison with state-of-the-arts: While our primary objective is to show the effectiveness of end-to-end person recognition over naeil with a similar setup, we provide comparisons with recent state-of-the-arts in Table 7.4 for completeness. We note that the key ingredients in each of these algorithms are differ in their choice of architecture, use of contextual information, external data and loss

Method	Accuracy $\mathcal{O}$	$\mathcal{A}$	$\mathcal{T}$	$\mathcal{D}$	Num of models	Architecture / parameters	Total param	Context used
Li *et al.* [74]	88.78	83.33	77.00	59.35	2	$\mathcal{DN}$ (15M) + $\mathcal{AN}$ (60M)	$\sim$75M	Yes
Li *et al.* [81]	84.93	78.25	66.43	43.73	2	VGG (130M)	$\sim$260M	Yes
PIPER [165]	83.05	-	-	-	107	$\mathcal{AN}$ (60M)	$\sim$6B	No
naeil [62]	86.78	78.72	69.29	46.61	17	$\mathcal{AN}$ (60M)	$\sim$1B	No
PSM [70]	89.21	82.73	74.84	56.73	16	$\mathcal{AN}$ (60M)	$\sim$1B	No
Liu *et al.* [88]	92.78	83.53	77.68	61.73	4	$\mathcal{IN}$ (40M*) + $\mathcal{RN}$ (45M)	$\sim$160M	No
N2NPR	**87.23**	**80.98**	**71.52**	**50.23**	1	$\mathcal{AN}$ (60M)	$\sim$60M	No

Table 7.4 Detailed comparison with previous recognition algorithms using PIPA dataset. M=millions, B=Billions. *Inception with extra fully connected layer having 87021 outputs).

Method	Train time (hrs)	Train speedup	Test* time (ms)	Test speedup
PIPER [165]	642	64x	160	32x
naeil [62]	102	10x	25	5x
PSM [70]	84	8x	20	4x
Liu *et al.* [88]	40	4x	60	12x
N2NPR	10	1x	5	1x

Table 7.5 Run time comparison of various approaches. *Time measured on a GTX 1080 Ti GPU with 10 CPUs

function. As a results, we mention system details of various algorithm to make comparisons more meaningful.

Following observations can be made from the table. Approaches [74, 81] that use contextual information achieve good performance even with two models. Without contextual information, approaches [62, 70, 165] using simpler architectures such as AlexNet require large number of models (in the range 8 to 107). Finally, with complex architecture (such as ResNet) and using region alignment procedure, top performance [88] can be achieved.

The performance of N2NPR is higher compared to naeil and comparable to current state-of-the-art PSM [70] that uses eight AlexNet models. This clearly demonstrates the usefulness of our approach. In [88], a combination of sophisticated models such as Inception and ResNet with better loss function leads to significantly improved performance. It should be noted that our approach is complementary to these methods and can leverage such techniques to achieve similar improvements. Finally, N2NPR

Method	Accuracy w/o tracks	Accuracy with tracks
Head (h)	27.52	31.91
Face (f)	26.53	31.55
Upper body (u)	16.49	17.72
Body (b)	14.76	16.53
h-f-u-b	29.18	34.72
naeil [62]	31.41	37.57
PSM [70]	40.95	44.46
N2NPR	**38.88**	**43.44**

Table 7.6 Overall frame and track level recognition performance computed using majority voting (%) on Hannah movie dataset. N2NPR obtains similar performance as that of PSM which requires large ensemble of models.

Method	Accuracy w/o tracks	Accuracy with tracks
Head (h)	17.68	20.54
Face (f)	14.32	17.35
Upper body (u)	17.62	20.68
Body (b)	17.51	20.32
h-f-u-b	18.12	21.59
naeil [62]	20.15	23.77
PSM [70]	20.48	24.31
N2NPR	**22.01**	**25.36**

Table 7.7 Recognition performance (%) of various approaches on the Soccer dataset. As the dataset contains significant face occlusions, our approach is able to shift the attention to the discriminative body regions and improve the state-of-the-art (PSM).

requires far less parameters compared to all the previous approaches, and hence is most suitable for practical applications.

7.4.3 Training and Testing Time

N2NPR requires significantly lower training and testing times compared to previous approaches. The is primarily due to reduced number of forward and backward operations required. Table 7.5 compares training and testing times of N2NPR with other approaches. For all the other approaches which require model ensemble training, we computed the total time based on time required for training/testing each region. For previous approaches which use images of size 224×224, it takes ~ 1.5 milli-seconds (ms)

for one image patch with AlexNet architecture while our input image 900×450 takes $\sim$4.8ms for four body regions for one forward pass. Our approach which benefits from computations sharing across regions runs much faster than the previous approaches, in both training and testing stages.

7.4.4 Results on Video Datasets

We now show results on newly introduced Hannah movie and soccer datasets for which we make comparison with baselines of `naeil` and `PSM` as provided in [70]. Results are provided in Table 7.6 and Table 7.7 for movie and soccer dataset, respectively at both frame and track level. We observe a significant improvement over `naeil` on both these datasets. Compared to the recent state-of-the-art `PSM` [70], we obtain a comparable performance on the movie dataset and outperform `PSM` achieving state-of-the-art results on soccer datasets. In challenging scenario of soccer, our approach based on adaptive weights is found to be more robust than `PSM` in handling large body deformations that are dominant in sport videos.

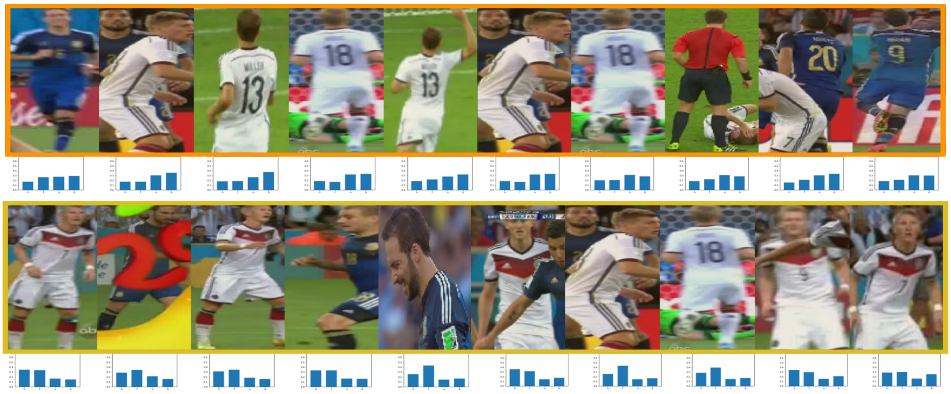
7.4.5 Qualitative results

Visualization of adaptive weights: In Figure 7.5, we show more examples of adaptive weights obtained on various test sets. It can be seen that head receives more weightage when it is completely or partially visible. Similarly, upper body receives more weightage for instances where head is not informative or when the person is facing away from the camera. Since head and face regions contain large overlapping regions, we notice similar weights between them. Similar is the case with upper body and body regions.

It is also interesting to look at the effectiveness of our adaptive weights in understanding different attributes of a given person image. As the weights indicate the discriminative ability of different body regions, images that have similar weights for body parts tend to have similar attributes in terms of visibility and image quality of the body parts. As shown in Figure 7.6, a clustering performed on the weight vectors result in images that have similar attributes of visibility and image quality. The images in the fourth cluster has their lower body occluded, which leads to high overlap of upper body and whole body regions.

(a) PIPA day test split

(b) Hannah movie dataset



(c) Soccer dataset

Figure 7.5 Adaptive weights obtained on (a) day split of PIPA (b) movie and (c) soccer test sets. Orange and yellow boxes indicate the scenarios where upper body (full body) and head (face) regions obtain high weightage, respectively when computing the final person representation. Notice how weights correspond to visibility and quality of the body region.

Figure 7.6 Clustering of attention weights tend to group images with similar discriminative parts together. This would be interesting in context of unsupervised attribute discovery.

Success and failure results on PIPA and soccer datasets: We show success and failure cases of our algorithm on PIPA dataset in Figures 7.7 and 7.8, respectively. Our algorithm is able to correctly predict the identities in diverse album scenarios with varied background, lighting, clothing and pose as shown in Figure 7.7. The failure results shown in Figure 7.8 demonstrate the negative influence of hair and clothing cues which cause misclassification in certain cases.

Similarly, we show such success and failure results on soccer dataset in Figure 7.9 and Figure 7.10, respectively. The soccer dataset is more challenging due to severe body deformations, occlusion, blur and non-visibility of regions. The nearest gallery images obtained for a probe image tend to have similar body pose or deformation. This suggests that certain kind of pose normalization or alignment may be required to identify people in sports domain.

Role of body regions: The role of head and upper body regions where their contribution resulted in correct and wrong predictions are shown in Figure 7.11 and Figure 7.12, respectively. Whenever head is clearly visible and has more weightage compared to upper body regions, it resulted in correct prediction (Figure 7.11). At the same time, when the head is occluded due to mask, hat or other accessories, it resulted in incorrect prediction as shown in Figure 7.12. We notice similar results for upper body region.

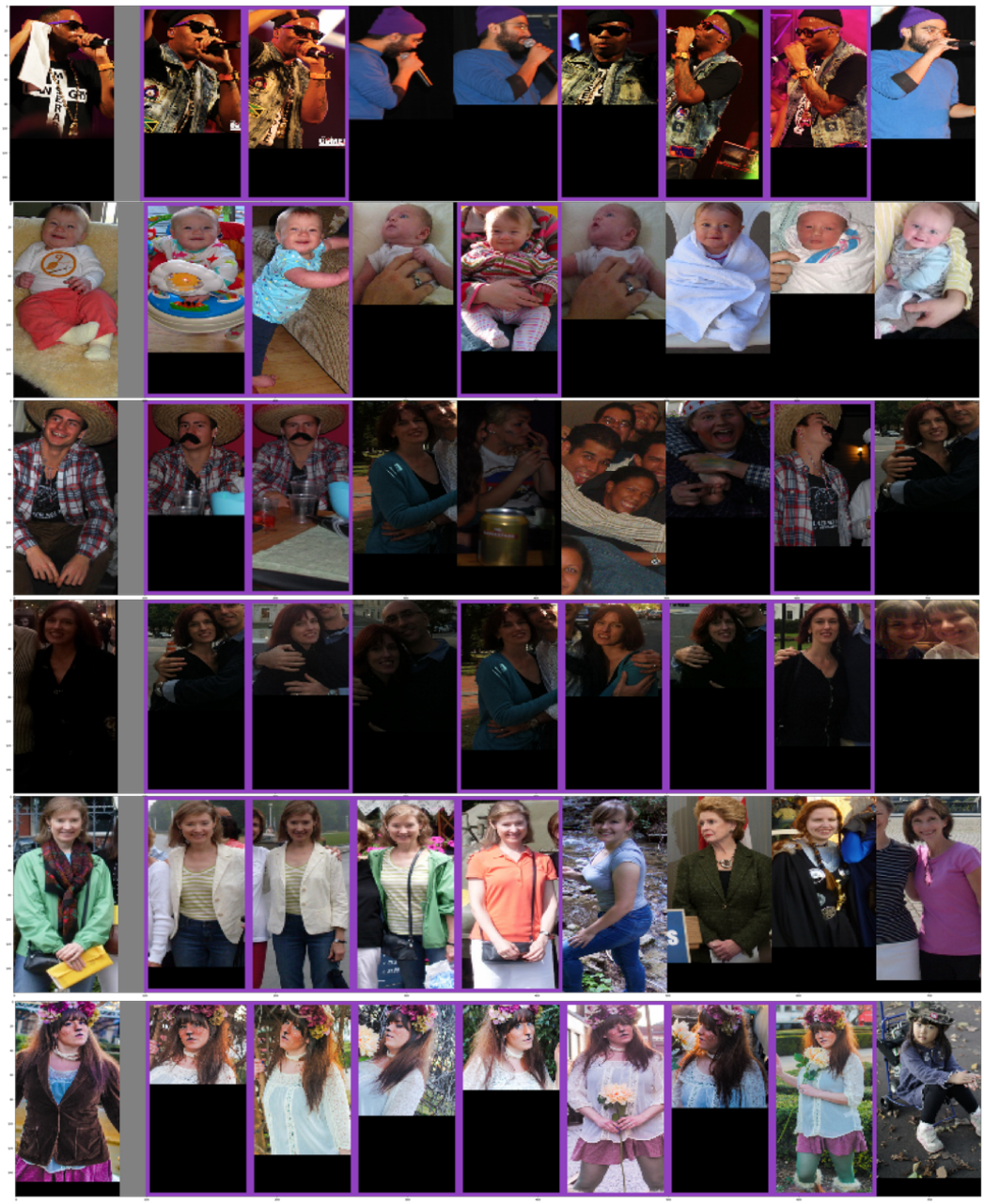


Figure 7.7 Success cases on PIPA day split. Each row shows a probe image (leftmost) and the gallery images that are closest to the probe image computed using our proposed person representation. Images shown inside the boxes are the neighbors from the correct class.

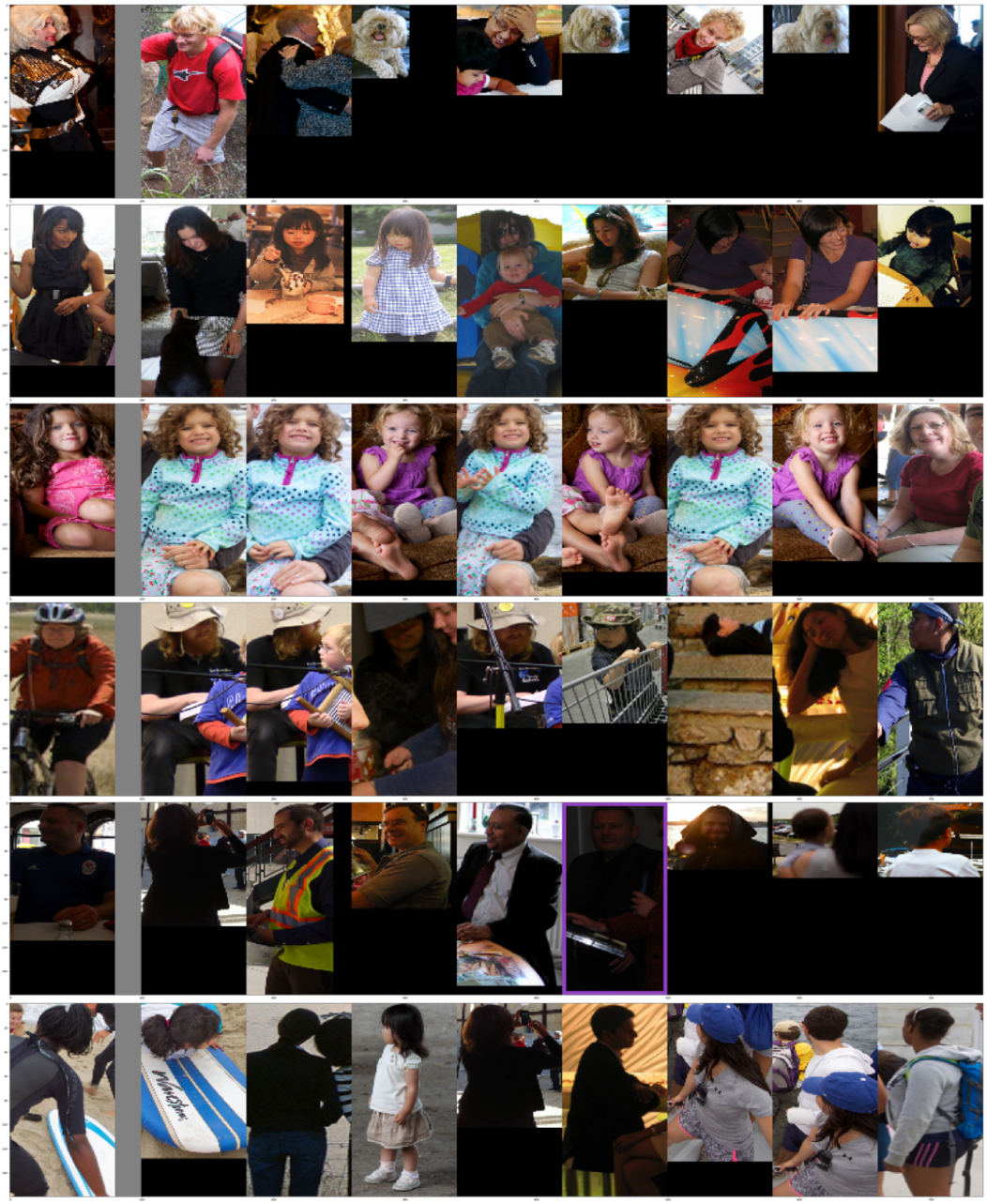


Figure 7.8 Failure cases on PIPA day split. Each rows show a probe image (leftmost) and the gallery images that are closest to the probe image computed using our proposed person representation. Images shown inside the boxes are the neighbors from the correct class. Notice how additional cues such as hair (first and third row), clothing (second and last row) can sometimes cause confusions.

Figure 7.9 Success cases on soccer dataset. Each row shows a probe image (leftmost) and the gallery images that are closest to the probe image computed using our proposed person representation. Images shown inside the boxes are the neighbors from the correct class.

Figure 7.10 Failure cases on soccer dataset. Each row shows a probe image (leftmost) and the gallery images that are closest to the probe image computed using our proposed person representation. Images shown inside the boxes are the neighbors from the correct class. For soccer scenario, in addition to blur, occlusion and non-visibility of regions which degrade the performance, we also see that the pose of the person and clothing can affect the prediction negatively, and it is necessary to develop more sophisticated approaches that account for these challenges.

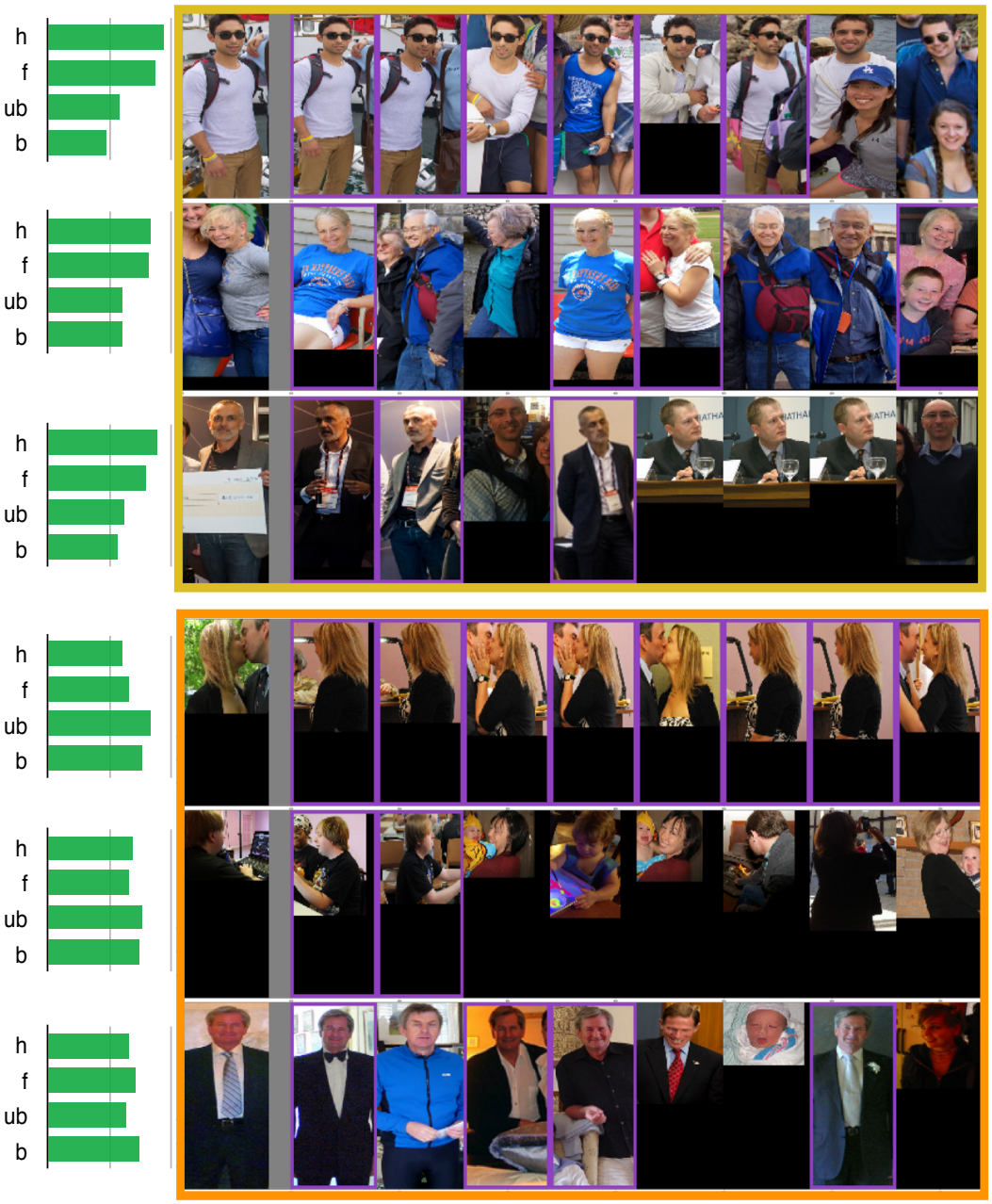


Figure 7.11 Success cases of head and upper body regions on PIPA day split. Each row shows the success predictions of our approach where the improvement is obtained primarily due to head /face (top three rows) or upper body/body (bottom three rows). The image in the left is the probe image and remaining images are the nearest gallery images based on our person representation. The weights on the left indicate the average weights of the nearest gallery images. Images shown inside the purple bounding boxes are the gallery images from the right class.

Figure 7.12 Failure cases of head and upper body regions on PIPA day split. Each row shows a failure prediction due to head /face (top three rows) and upper body/body (bottom three rows). The image in the left is the probe image and remaining images are the nearest gallery images based on our person representation. The weights on the left indicate the average weights of the nearest gallery images. Images shown inside the purple bounding boxes are the gallery images from the right class.

7.5 Summary

We introduced a novel end-to-end approach for person recognition based on region pooling and attention mechanisms. Our framework employs region pooling to pool regional features from person feature maps. An adaptive weighting scheme based on attention mechanism then combines these regional features to produce final representation. Our end-to-end scheme achieves results that are comparable with previous approaches on PIPA and movie dataset, and produces state-of-the-art results on soccer dataset. The approach is simpler, faster and requires far less parameters and hence is suited for practical applications.

Chapter 8

Summary, Conclusion and Scope of Future Work

Here, we conclude by summarizing our thesis and highlighting some of the future research directions.

In our thesis, we focused on person recognition from the practical standpoint and looked into different sub-problems seen in practical settings. Specifically, we considered the issues of availability of training samples, domain mismatch between instances, and non-visibility of faces. In all these scenarios, our proposed solutions exploit additional or complimentary information to improve the performance. While similarities with unlabeled examples are exploited when the training examples are limited, temporal and appearance similarities are exploited in video or image collections. And for the cases where face is not completely visible, cues from multiple body regions are used to improve the performance.

In the beginning of our thesis, we provided the motivation for the problem of person recognition, its applications and various challenges involved in the practical settings, and provided a survey of related person detection and recognition techniques. We first looked into person detection in chapter 3 which forms the critical precursor to any recognition problem. Our work was built upon the ideas of exemplar based face detection. While the exemplar detector, having derived originally from retrieval framework, has many benefits such as non-sliding window scanning, scalability, and flexibility to retrain, it had limited performance due to less discriminative scoring mechanism and ignorance of visual feature dependencies. We introduced the notion of visual phrases and spatial context into the exemplar framework to overcome these issues, thereby improving its performance significantly.

In the subsequent chapters, we focused mainly on recognition, but in several practical settings. We considered the limitations of training data in chapter 4 and demonstrated how unlabeled data can be leveraged in such scenario. Our proposed technique based on a iterative algorithm that selectively labels unlabeled examples in a successive fashion proved to improve the recognition significantly. We then considered the issue of domain mismatch in chapter 5. We considered image collection as our

experimental setup where the objective is to recognize images in a test collection using training images from completely different domain. We showed the effectiveness of two-stage approach for the task. While first stage focuses on labeled seed image generation from the target domain, the second stage focuses on propagating the labels from seed images to collection images. While exemplar framework is adopted for detection task, sparse representation classifier and label propagation algorithms are applied for recognition.

In the last two chapters, we went beyond faces for recognition and explored the effectiveness of different body regions such as head, upper body, *etc*, for recognition. We proposed a person recognition approach that tackles pose variations by training an ensemble of pose-specific models, each trained on head and upper body regions in a particular pose. A deep neural network that jointly optimizes the objective function for multiple body regions is also proposed to learn pose-specific person representation. While multi-region models seem to be effective for person recognition, training multiple deep neural networks for each region is found to be a time consuming process, requiring large number of parameters and memory. To overcome these issues, we developed an end-to-end person recognition model in the final chapter that requires a single model with far less parameters and is ideal for practical applications. An end-to-end model that takes complete person image as input and generates a compact person representation is proposed. We finally introduced two person benchmarks based on movie and soccer videos and made it publicly available.

8.1 Conclusion

World around us is changing at a rapid pace and many tasks that require significant human effort are being automated. Thanks to several break though results in artificial intelligence (AI) in the last few years, we are witnessing a great progress and many of tasks that seemed impossible a few years ago are now getting deployed in the real world. There is a also a great buzz around AI and its possibilities, both in academic and commercial organizations.

Person recognition has witnessed a similar growth story in the recent years. The timing of our work could not have been better since facial and person recognition technologies are being deployed in variety of applications such as mobile device unlocking (iPhone), photo-organization (Google photos, Facebook photos), airport security authentication, traffic surveillance, etc. As our thesis looks from deployment

perspective and tackles several practical issues, we hope that thesis would benefit readers who wish to develop image based recognition systems.

8.2 Future Directions

In our thesis, we have proposed approaches for improving person recognition in various scenarios. Our future directions are focused on developing better features, datasets and algorithms as listed below.

- **Exemplar detection with deep features:** Exemplar approaches described in Chapter 3 use dense shallow features (SIFT) for face detection. Inspired by the success of CNN features for detection, an immediate possibility is to revisit the exemplar detector with CNN features. Since the requirement for exemplar detectors are local features and visual word indexing, the work requires exploring techniques to obtain local descriptors from CNN activation maps.

- **Person recognition at low resolution:** From our preliminary experiments on datasets such as soccer videos, we found that the performance of existing methods is very poor. One of the main issues is the scale of the objects. The object features that are discriminative at larger scales are completely different from those at small scales. As a result, it is necessary to develop new techniques that can recognize people at very small scale. This is also useful for surveillance and other security related applications.

- **Open-set recognition:** Many of the classification problems such as object recognition are closed-set problems where it is expected to predict same set of classes, both during training and testing. Softmax loss function is popularly applied in such case. However, person recognition is inherently an open-set problem where the training subjects are different compared to gallery and probe subjects. It would therefore be effective to develop open-set compatible loss functions such that networks trained on one set of subjects are discriminative enough to match different set of subjects between gallery and probe images.

- **Unsupervised feature learning:** The CNN representation learning for faces require large labeled training datasets. A question that is worth studying in the context of person recognition is the performance of unsupervised features obtained through generative models such as auto-encoders or generative adversarial networks (GAN).

- **Larger datasets and better algorithms:** There is a large gap between the real-world performance of current algorithms compared to performance on existing datasets. As seen in Table 2.1, Table 2.2 and Table 2.3, existing detection and recognition benchmarks are *saturated* as the algorithms started producing near 100% accuracy. However, the same performance cannot be guaranteed in the real-world. One of the future possibilities is to reduce this gap by creating larger and complex datasets that facilitate future research in person recognition domain.

Bibliography

[1] Apriori. `http://www.borgelt.net/apriori.html`.

[2] Dlib C++ Library. `http://dlib.net/`.

[3] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan, et al. Face recognition using deep multi-pose representations. In *IEEE Winter Conference on Computer Vision*, 2016.

[4] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *International Conference on Very Large Scale Data Bases*, 1994.

[5] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[6] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[7] D. Anguelov, K.-c. Lee, S. B. Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[8] A. Babenko and V. Lempitsky. Aggregating local deep features for image retrieval. In *IEEE international conference on computer vision*, 2015.

[9] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 2003.

[10] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 2006.

[11] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? *arXiv preprint arXiv:1411.4304*, 2014.

[12] T. Berg and P. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013.

[13] M. Bertini, A. Del Bimbo, and W. Nunziati. Player identification in soccer videos. In *SIGMM workshop on Multimedia information retrieval*, 2005.

[14] W. W. Bledsoe. The model method in facial recognition. In *Panoramic Research Inc., Palo Alto, CA*, 1966.

[15] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *IEEE International Conference on Computer Vision*, 2009.

[16] M. Buml, M. Tapaswi, and R. Stiefelhagen. Semi-supervised learning with constraints for person identification in multimedia data. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013.

[17] D. Cai, X. He, and J. Han. Semi-supervised discriminant analysis. In *IEEE International Conference on Computer Vision*, 2007.

[18] O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 2009.

[19] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision*, 2014.

[20] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *British Machine Vision Conference*, 2011.

[21] G. Chéron, I. Laptev, and C. Schmid. P-CNN: Pose-based CNN Features for Action Recognition. In *IEEE International Conference on Computer Vision*, 2015.

[22] O. Chum and J. Matas. Unsupervised discovery of co-occurrence in sparse high dimensional data. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[23] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2005.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[25] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Pedestrian attribute recognition at far distance. In *IEEE MultiMedia*, 2014.

[26] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[27] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. *arXiv preprint arXiv:1704.05020*, 2017.

[28] G. Edwards, C. Taylor, and T. Cootes. Improving identification performance by integrating evidence from sequences. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1999.

[29] M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automated naming of characters in TV video. *Image and Vision Computing*, 2009.

[30] H. Fan, H. Su, and L. Guibas. A point set generation network for 3d object reconstruction from a single image. *arXiv preprint arXiv:1612.00603*, 2016.

[31] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[32] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[33] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[34] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 1973.

[35] S. C. Fralick. Learning to recognize patterns without a teacher. *IEEE Transactions on Information Theory*, 1967.

[36] J. Gall, A. Yao, N. Razavi, L. J. V. Gool, and V. S. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[37] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[38] V. Gandhi and R. Ronfard. Detecting and naming actors in movies using generative appearance models. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013.

[39] R. Garg, S. M. Seitz, D. Ramanan, and N. Snavely. Where's waldo: Matching people in images of crowds. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[40] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.

[41] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015.

[42] R. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, 2015.

[43] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE International Conference Computer Vision and Pattern Recognition*, 2014.

[44] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*. Springer, 2014.

[45] D. Gorodnichy. On Importance of Nose for Face Tracking. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[46] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *European Conference on Computer Vision*, 2008.

[47] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *IEEE International Conference on Computer Vision*, 2009.

[48] S. H and K. T. Object detection using the statistics of parts. *International Journal of Computer Vision*, 2004.

[49] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[50] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, 2017.

[51] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *IEEE European Conference on Computer Vision*, 2014.

[52] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[53] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *IEEE European Conference on Computer Vision*, 2012.

[54] P. Hu and D. Ramanan. Tiny Faces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[55] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

[56] V. Jain and E. Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[57] V. Jain and E. Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[58] H. Jégou and O. Chum. Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. In *IEEE European Conference on Computer Vision*, 2012.

[59] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[60] H. Jiang and E. Learned-Miller. Face detection with the faster r-cnn. *arXiv preprint arXiv:1606.03473*, 2016.

[61] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, 2010.

[62] S. Joon Oh, R. Benenson, M. Fritz, and B. Schiele. Person recognition in personal photo collections. In *IEEE International Conference on Computer Vision*, 2015.

[63] T. Kanade. Picture processing system by computer complex and recognition of human faces. In *Dept. of Information Science, Kyoto University*, 1973.

[64] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale l_1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, 2007.

[65] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[66] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[67] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, 2009.

[68] V. Kumar, A. Namboodiri, and C V Jawahar. Sparse representation based face recognition with limited labeled samples. In *Asian Conference on Pattern Recognition*, 2013.

[69] V. Kumar, A. Namboodiri, and C V Jawahar. Visual phrases for exemplar face detection. In *IEEE International Conference on Computer Vision*, 2015.

[70] V. Kumar, A. Namboodiri, M. Paluri, and C. Jawahar. Pose-aware person recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017.

[71] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back. Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 1997.

[72] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[73] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *IEEE European Conference on Computer Vision, workshop on statistical learning in computer vision*, 2004.

[74] H. Li, J. Brandt, Z. Lin, X. Shen, and G. Hua. A multi-level contextual model for person recognition in photo albums. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[75] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In *IEEE International Conference on Computer Vision*, 2013.

[76] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua. Efficient boosted exemplar-based face detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[77] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[78] J. Li and Y. Zhang. Learning SURF cascade for fast and accurate object detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013.

[79] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[80] W. Li, R. Zhao, T. Xiao, and X. Wang. DeepReId: Deep filter pairing neural network for person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[81] Y. Li, G. Lin, B. Zhuang, L. Liu, C. Shen, and A. v. d. Hengel. Sequential person recognition in photo albums with a recurrent network. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017.

[82] Y. Li, B. Sun, T. Wu, and Y. Wang. Face detection with end-to-end integration of a convnet and a 3D model. In *European Conference on Computer Vision*, 2016.

[83] S. Liao, A. K. Jain, and S. Z. Li. A fast and accurate unconstrained face detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[84] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Joint Pattern Recognition Symposium*, 2003.

[85] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear CNN models for fine-grained visual recognition. In *IEEE International Conference on Computer Vision*, 2015.

[86] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *IEEE European Conference on Computer Vision*, 2012.

[87] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *European conference on computer vision*, 2016.

[88] Y. Liu, H. Li, and X. Wang. Rethinking feature discrimination and polymerization for large-scale recognition. In *Advances in Neural Information Processing Systems*, 2017.

[89] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017.

[90] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy. Learning to track and identify players from broadcast sports videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[91] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *IEEE European Conference on Computer Vision*, 2012.

[92] J. Malik, P. Arbeláez, J. Carreira, K. Fragkiadaki, R. Girshick, G. Gkioxari, S. Gupta, B. Hariharan, A. Kar, and S. Tulsiani. The three R's of computer vision: Recognition, Reconstruction and Reorganization. *PR Letters*, 2016.

[93] N. Markuš, M. Frljak, I. S. Pandzic, J. Ahlberg, and R. Forchheimer. A method for object detection based on pixel intensity comparisons. In *Second Croatian Computer Vision Workshop*, 2013.

[94] A. Martinez and R. Benavente. The AR face database. CVC Technical Report #24, June 1998.

[95] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[96] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *IEEE European Conference on Computer Vision*, 2014.

[97] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Applications*, 2009.

[98] A. Nech and I. Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017.

[99] J. Y.-H. Ng, F. Yang, and L. S. Davis. Exploiting local features from deep networks for image retrieval. 2015.

[100] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 2000.

[101] E. Ortiz, A. Wright, and M. Shah. Face Recognition in Movie Trailers via Mean Sequence Sparse representation-based classification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013.

[102] A. Ozerov, J.-R. Vigouroux, L. Chevallier, and P. Perez. Clothing cosegmentation for recognizing people. In *International Conference on Image Processing*, 2013.

[103] F. P and H. D. Pictorial structures for object recognition. *International Journal of Computer Vision*, 2005.

[104] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. *British Machine Vision Conference*, 2015.

[105] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 2015.

[106] N. Pinto, J. J. DiCarlo, and D. D. Cox. How far can you get with a modern face recognition test set using only simple features? In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[107] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *ACM International Conference on Machine Learning*, 2007.

[108] R. Ranjan, V. M. Patel, and R. Chellappa. A deep pyramid deformable part model for face detection. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, 2015.

[109] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv:1603.01249*, 2016.

[110] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *IEEE International Conference Computer Vision and Pattern Recognition*, 2016.

[111] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[112] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

[113] F. Roli and G. L. Marcialis. Semi-supervised pca-based face recognition using self training. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition*, 2006.

[114] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[115] T. Sakai, M. Nagao, and S. Fujibayashi. Line Extraction and Pattern Detection in a Photograph. *Pattern Recognition*, 1969.

[116] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[117] H. J. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 1965.

[118] N. Sebe, I. Cohen, T. S. Huang, and T. Gevers. Semi-supervised face detection. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2005.

[119] S. Setty, M. Husain, P. Beham, J. Gudavalli, M. Kandasamy, R. Vaddi, V. Hemadri, J. C. Karure, R. Raju, Rajan, V. Kumar, and C. V. Jawahar. Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations. In *National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, 2013.

[120] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *IEEE European Conference on Computer Vision*, 2002.

[121] C. Shan. Face recognition and retrieval in video. In *Video Search and Mining*, 2010.

[122] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012.

[123] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Detecting and aligning faces by image retrieval. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013.

[124] M. Shi, X. Sun, D. Tao, and C. Xu. Exploiting visual word co-occurrence for image retrieval. In *IEEE Multimedia*, 2012.

[125] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[126] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *British Machine Vision Conference*, 2013.

[127] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[128] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" Learning person specific classifiers from video. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[129] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, 2003.

[130] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *British Machine Vision Conference*, 2006.

[131] S. Sukhbaatar, J. Weston, R. Fergus, et al. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, 2015.

[132] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Neural Information Processing Systems*, 2014.

[133] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[134] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[135] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[136] M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Knock! Knock! Who is it? a probabilistic person identification in TV-series. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012.

[137] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 1991.

[138] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013.

[139] E. Ustinova, Y. Ganin, and V. Lempitsky. Multiregion bilinear convolutional neural networks for person re-identification. *arXiv preprint arXiv:1512.05300*, 2015.

[140] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. 2010.

[141] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.

[142] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 2012.

[143] T. Vu, A. Osokin, and I. Laptev. Context-aware CNNs for person head detection. In *IEEE International Conference on Computer Vision*, 2015.

[144] F. Wang and C. Zhang. Label propagation through linear neighborhoods. *IEEE Transactions on Knowledge and Data Engineering*, 2008.

[145] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[146] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han. Contextual weighting for vocabulary tree based image retrieval. In *IEEE International Conference on Computer Vision*, 2011.

[147] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li. Detecting faces using region-based fully convolutional networks. *arXiv preprint arXiv:1709.05256*, 2017.

[148] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[149] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Workshop on faces in 'real-life' images: Detection, alignment, and recognition*, 2008.

[150] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[151] J. Yan, Z. Lei, L. Wen, and S. Z. Li. The fastest deformable part model for object detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[152] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image and Vision Computing*, 2014.

[153] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Aggregate channel features for multi-view face detection. In *IEEE International Joint Conference on Biometerics*, 2014.

[154] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017.

[155] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[156] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[157] R. B. Yeh, A. Paepcke, H. Garcia-Molina, and M. Naaman. Leveraging context to resolve identity in photo albums. In *Proceedings of the 5th Joint Conference on Digital Libraries*, 2005.

[158] D. Yi, Z. Lei, and S. Z. Li. Towards pose robust face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013.

[159] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *International Conference on Pattern Recognition*, 2014.

[160] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[161] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[162] L. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? *IEEE International Conference on Computer Vision*, 2011.

[163] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for sub-category recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012.

[164] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[165] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[166] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[167] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *ACM Multi Media*, 2009.

[168] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry-preserving visual phrases. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[169] M. Zhao, J. Yagnik, H. Adam, and D. Bau. Large scale learning and recognition of faces in web videos. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[170] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[171] W.-Y. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 2003.

[172] X. Zhao, N. W. Evans, and J.-L. Dugelay. Semi-supervised face recognition with LDA self-training. In *IEEE International Conference on Image Processing*, 2011.

[173] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. *Advances in Neural Information Processing Systems*, 2004.

[174] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *arXiv preprint arXiv:1501.04690*, 2015.

[175] X. Zhu. Semi-supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison, 2005.

[176] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Carnegie Mellon University, 2002.

[177] X. Zhu and A. B. Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 2009.

[178] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[179] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE International Conference Computer Vision and Pattern Recognition*, 2012.

[180] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543*, 2014.