**ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY**

# Face Track Retrieval and Recognition Across Age

**M.Sc. THESIS**

**Esam Ghaleb**

**Department of Computer Engineering**

**Computer Engineering Programme**

**August 2015**

**ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY**

# Face Track Retrieval and Recognition Across Age

**M.Sc. THESIS**

**Esam Ghaleb**
**(504131515)**

**Department of Computer Engineering**

**Computer Engineering Programme**

**Thesis Advisor: Assoc. Prof. Dr. Hazım Kemal EKENEL**

**August 2015**

**İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ**

# Yaşlar Arası Yüz İz Çıkarımı ve Tanıması

**YÜKSEK LİSANS TEZİ**

**Esam Ghaleb**
**(504131515)**

**Bilgisayar Mühendisliği Anabilim Dalı**

**Bilgisayar Mühendisliği Programı**

**Tez Danışmanı: Assoc. Prof. Dr. Hazım Kemal EKENEL**

**Ağustos 2015**

**Esam Ghaleb**, a M.Sc. student of ITU Graduate School of Science Engineering and Technology 504131515 successfully defended the thesis entitled **"Face Track Retrieval and Recognition Across Age"**, which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

| | | |
|---|---|---|
| **Thesis Advisor :** | **Assoc. Prof. Dr. Hazım Kemal EKENEL**<br>Istanbul Technical University | ..... |
| **Jury Members :** | **Prof. Dr. Muhittin Gökmen**<br>MEF University | ..... |
| | **Prof. Dr. Zehra Çataltepe**<br>Istanbul Technical University | ..... |
| | | ..... |

**Date of Submission :** **17 August 2015**
**Date of Defense :** **21 August 2015**

***To my family,***

## FOREWORD

Video face recognition is a very popular task and has come a long way. The primary challenges such as illumination, resolution and pose are well studied through multiple datasets and have extensive researches. However, age invariant face recognition still lacks enough studies and there are no video-based datasets dedicated to study the effects of aging on facial appearance. Unlike previous works that deal with studying age progression using still images, our study focuses on face track retrieval and recognition across age using video based dataset. We present a challenging face track dataset, Harry Potter Movies Aging dataset (Accio), to study and develop age invariant face recognition methods for videos. In addition we conduct an extensive study on the dataset for purpose of proving the aging factor in the dataset and for obtaining an age invariant face track retrieval and recognition. Our dataset not only has strong challenges of pose, illumination and distractors, but also spans a period of ten years providing substantial variation in facial appearance. It consists of large number of face tracks for variant characters and age groups. It provides a good environment and case to study the impact of age progression together with other technical challenges in face recognition in videos such face pose variation, expression, illumination and resolution. We make the dataset publicly available for further exploration in age-invariant video face recognition.

We propose two primary tasks: within and across movies face track retrieval and recognition; and three protocols which differ in their freedom to use Accio or external data. We achieve retrieval experiments, which are evaluated using two popular measures: (i) mean average precision (mAP), and (ii) precision @ $k$. We present baseline and improved results for the retrieval using a state-of-the-art face track descriptor (Fisher vector). Our baseline experiments show clear trends of reduction in performance as the age gap between the query or test and database or train movies increases. In order to improve the performance of baseline: we adopt three metric learning techniques: (i) *joint metric learning*, (ii) *low-rank Mahalanobis metric learning*, and (iii) *diagonal metric learning*. We use these metric learning algorithms in two different scenarios: (i) metric learning on pairwise movies, which obtain a metric model from a percentage of face tracks of two movies and (ii) metric learning on all movies which combine a percentage of face tracks from all movies. In both scenarios, we observe an improvement over the baseline results in case of using joint and low-rank Mahalanobis metric learning in term of both mAP and precision @k. They reduce the gap between cross movies' results that emerge due to facial changes across movies and they increase the results of mAP and precision @k. The best average improvement compared to the baseline is obtained in case of using joint metric learned from all movies which reach to 42.2%. Most of improvements are between movies with large gap between their release date. This due to the ability of the used metric learning to capture the similarity features across ages.

In addition, we apply linear Support Vector Machines (SVMs) on the dataset for both evaluation tasks in order to assess of the performance of face track recognition across age on this dataset. The accuracy results of this experiment support the trend we obtain

in retrieval evaluation by giving results in similar similar behavior. We always obtain higher recognition rate from the evaluation within the same movies, however the rate decreases as we test across movies especially when the years' gap increases.

As a part of our aforementioned study, we measure the performance of the state-of-the-art face verification method: Fisher vector. We carry out several experiments on Fisher vector using different dataset, Facial Recognition Grand Challenge (FRGC), rather than Label Face in the Wild (LFW). In addition we benefit from Gabor filters as local descriptor rather than Scale Invariant Feature Transformation(SIFT). In addition we use various settings of data and Fisher vector such as, face alignment, Gaussian Mixture Model (GMM) size, Principal Component Analysis ( PCA) dimensions on features, and the information augmentation local descriptors. We evaluate each of these parameters separately in order to check their impact on the Fisher vectors on face verification task. In this work we provide an expanded analysis of each parameter with comparison to different setting.

The major contributions of the this work can be summarized as following: we present "Harry Potter Movies Aging Dataset (Accio)", provide a baseline and improved face track retrieval results using Fisher vector face track descriptor and metric learning techniques. We apply face track recognition evaluation which also endorse the retrieval baseline results to prove aging factor in Accio dataset. Finally we carry out Fisher vector analysis to evaluate its performance on face verification.

**ACKNOWLEDGMENT**

I would like to express my deepest expression and thanks for everyone who helped me to complete this thesis. First, I would like to express my gratitude to Assoc. Prof. Hazım Kemal Ekenel for his supervision of my thesis, patience, and providing me with a great atmosphere for doing research in Smart Interaction, Mobile Intelligence, and Multimedia Technologies (SiMiT) Research Group. I also thank everyone in SiMiT Lab for their support and having great time with them during my work.

I would like to thank Prof. Rainer Stiefelhagen for providing me an opportunity to work on my thesis during my summer visits to Computer Vision for Human Computer Interaction (CVHCI) Laboratory in Karlsruhe Institute of Technology (KIT). It had been a great experience with the inspiring research group in CVHCI. I would like to thank all the members of the laboratory who made my stay fantastic.

Very special thanks goes to Makarand Tapaswi, for being excellent advisor and for his untiring help. I am very grateful to have him as an advisor. I thank him for suggesting this exciting work of my thesis. It was his idea to work on face track retrieval and recognition across age using challenging and novel face track dataset called "Harry Potter Movies Aging Dataset (Accio)". I also thank him for helping me to prepare and release the dataset. I appreciate his vast knowledge and skills in many areas ( e.g., guidance, vision, research), sharing his experience, and his assistance in writing and developing this work. Makarand Tapaswi is a person who truly made difference in my life. Without his motivation and encouragement I would not have complete this work.

I am also grateful for Ziad Al-Halah for his suggestions and directions during my work. He always asks the right questions and we have interesting discussions during our meet with Makarand Tapaswi.

In addition, this work is funded by the German Research Foundation under contract no. STI-598/2-1; TUBITAK within the CHIST-ERA project titled CAMOMILE, project no. 112E176; and a Marie Curie FP7 Integration Grant in the 7th EU Framework.

Last but not least, I want to thank my family in Yemen: my parents, my sisters, and my brothers who always guide and support me spiritually to excel throughout my life.

Ağustos 2015 Esam Ghaleb

**TABLE OF CONTENTS**

**Page**

**FOREWORD** ..... **xi**
**TABLE OF CONTENTS** ..... **xiii**
**ABBREVIATIONS** ..... **xv**
**LIST OF TABLES** ..... **xvii**
**LIST OF FIGURES** ..... **xix**
**SUMMARY** ..... **xxi**
**ÖZET** ..... **xxv**
**1. INTRODUCTION** ..... **1**
1.1 Motivation and Hypothesis ..... 4
1.2 Problem Definition ..... 6
1.3 Thesis Outline ..... 6
**2. RELATED WORK** ..... **9**
2.1 Face Identification ..... 9
2.2 Data Sets ..... 10
2.3 Face Track Retrieval ..... 10
2.4 Age Invariant Face Recognition ..... 11
**3. THEORETICAL FOUNDATION** ..... **13**
3.1 Shot Boundary Detection ..... 13
3.2 Face Detection and Tracking ..... 14
3.2.1 Modified census transformation ..... 14
3.2.2 Boosting and cascade classifiers ..... 15
3.2.3 Face tracking ..... 15
3.2.4 Face alignment ..... 16
3.3 Face Image and Track Representation ..... 17
3.3.1 Fisher vector for face recognition ..... 18
3.4 Metric Learning ..... 21
3.4.1 Large margin dimensionality reduction ..... 21
3.4.2 Low-rank joint metric learning ..... 23
3.4.3 Diagonal metric learning ..... 24
3.5 Linear Support Vector Machines (SVMs) ..... 25
3.6 Performance Evaluation Methods ..... 26
3.6.1 Receiving operator characteristic (ROC) curve: ..... 27
**4. FACE TRACK RETRIEVAL AND RECOGNITION ACROSS AGE** ..... **29**
4.1 Accio: Harry Potter Movies Aging Dataset ..... 29
4.1.1 Data source ..... 29
4.1.2 Face track descriptors ..... 31
4.1.3 Statistics ..... 31

4.2 Experiments ..... 33
4.2.1 Evaluation protocol ..... 33
4.2.2 Baseline evaluation on all face tracks ..... 34
4.3 Improving Baseline Face Track Retrieval Results ..... 37
4.3.1 Metric learning on pairwise movies ..... 37
4.3.2 Metric learning on all movies data ..... 47
4.4 Face Track Recognition ..... 55
**5. FACE VERIFICATION ANALYSIS USING FISHER VECTORS ..... 59**
5.1 Face Recognition Grand Challenge(FRGC) Data Set ..... 59
5.2 Experiments: FRGC's Fourth Experiment ..... 60
5.2.1 SIFT features ..... 60
5.3 Fisher Vectors Using Gabor Filters ..... 61
5.3.1 Spatial, scale and orientations augmentations: ..... 62
5.4 Experiments and Results ..... 63
5.4.1 Effect of face alignment ..... 63
5.4.2 Effect of PCA dimension size ..... 64
5.4.3 Effect of spatial, scale and orientation augmentation ..... 64
5.4.4 Effect of GMM size: ..... 65
5.4.5 Final experiment ..... 66
**6. CONCLUSION AND FUTURE WORK ..... 69**
6.1 Conclusion ..... 69
6.2 Future Work ..... 70
**REFERENCES ..... 73**
**CURRICULUM VITAE ..... 79**

## ABBREVIATIONS

**mAP** : Mean Average Precision
**AP** : Average Precision
**SVMs** : Support Vector Machines
**FRGC** : Facial Grand Challenge
**SIFT** : Scale Invariant Feature Transformation
**GMM** : Gaussian Mixture Model
**PCA** : Principal Component Analysis
**MCT** : Modified Census Transformation
**LFW** : Label Face in The Wild
**DFD** : Displaced Frame Difference
**ROC** : Receiving Operator Curve
**SGD** : Stochastic Gradient Descent

## LIST OF TABLES

**Page**


**Table 4.1 :** Number of tracks and characters across the movies. ..... 32
**Table 4.2 :** Actors age distribution. ..... 32
**Table 4.3 :** Comparison of age-invariant face recognition datasets..... 33
**Table 4.4 :** Within movies evaluation. ..... 35
**Table 4.5 :** across movies face track retrieval performance (MAP). ..... 35
**Table 4.6 :** Comparison of MAP scores for adult vs. young actors. The queries are taken here from HP-1 movie. ..... 37
**Table 4.7 :** Comparison of number of pairs when using different percent of face tracks from pair movies. ..... 38
**Table 4.8 :** Pairwise baseline's cross movies FT retrieval: mAP. ..... 38
**Table 4.9 :** Pairwise joint metric's cross movies FT retrieval: mAP. ..... 40
**Table 4.10 :** The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values..... 40
**Table 4.11 :** Cross movies FT retrieval: mAP. ..... 42
**Table 4.12 :** Full dimensional Fisher vectors: The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values..... 43
**Table 4.13 :** Pairwise low rank Mahalanobis metric's cross movies FT retrieval: mAP. ..... 43
**Table 4.14 :** The average differences of the precision @k between baseline and low-rank Mahalanobis metric learning that show the improvement in tasks 1 and 2 results for different k values..... 44
**Table 4.15 :** Pairwise diagonal metric's cross movies FT retrieval: mAP. ..... 44
**Table 4.16 :** The average differences of the precision @k between baseline and diagonal metric learning that show the improvement in tasks 1 and 2 results for different k values..... 44
**Table 4.17 :** All movies baseline's cross movies FT retrieval: mAP. ..... 47
**Table 4.18 :** All movies joint metric's cross movies FT retrieval: mAP. ..... 48
**Table 4.19 :** Metric learning from from all movies: The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values. 49
**Table 4.20 :** Full dimensional Fisher vectors - All movies joint metric's cross movies FT retrieval: mAP. ..... 50
**Table 4.21 :** Metric learning from from all movies on full dimensional Fisher vectors: The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values. ..... 51

**Table 4.22 :** All movies low-rank Mahalanobis metric's cross movies FT retrieval: mAP. ..... 51
**Table 4.23 :** Metric learning from from all movies: The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values. 51
**Table 4.24 :** All movies diagonal metric's cross movies FT retrieval: mAP. ..... 53
**Table 4.25 :** Metric learning from from all movies: The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values. 54
**Table 4.26 :** SVM accuracy: overall. ..... 56
**Table 4.27 :** F1 score of face recognition for Harry Potter character in Accio dataset. ..... 57
**Table 4.28 :** F1 score of face recognition for Hermione character in Accio dataset. ..... 57
**Table 4.29 :** F1 score of face recognition for Ron Weasley character in Accio dataset. ..... 57
**Table 5.1 :** Evaluation results of LFW splits in term of ROC = ( 1 - EER). ..... 60
**Table 5.2 :** Using SIFT features: AUC and VR@0.1 FAR% of FRGC's fourth experiment..... 61
**Table 5.3 :** Effect of face alignment: AUC and VR@0.1% FAR of FRGC's fourth experiment..... 64
**Table 5.4 :** Effect of PCA Dimensions AUC and VR@0.1% FAR of FRGC's fourth experiment..... 64
**Table 5.5 :** Effect of Augmentation: AUC and VR@FAR = 0.1% FAR of FRGC's fourth experiment..... 65
**Table 5.6 :** Effect of GMM size: AUC and VR@0.1% FAR of FRGC's fourth experiment..... 66
**Table 5.7 :** Final Results of AUC and VR@0.1% FAR of FRGC's fourth experiment..... 66
**Table 5.8 :** Comparison between SIFT and Gabor results of AUC and VR@0.1% FAR of FRGC's fourth experiment. ..... 68

## LIST OF FIGURES

**Page**


**Figure 1.1** : Change in character appearance across the movies. ..... 2
**Figure 1.2** : Across movies evaluation hypothesis: Dominant matrix. ..... 5
**Figure 3.1** : Shot boundary example [1]..... 13
**Figure 3.2** : Face detection with MCT features and cascade classifiers [2]. ..... 15
**Figure 3.3** : Face tracks obtained by face tracker in Harry Potter movie series.... 17
**Figure 3.4** : Aligned faces with different eye distances and eye rows. ..... 18
**Figure 3.5** : Fisher vector implementation pipeline [3]..... 18
**Figure 3.6** : Gabor wavelets with 8 orientations and 5 scales. ..... 20
**Figure 3.7** : Positive and negative pairs are separated with a margin by metric learning [3]..... 22
**Figure 4.1** : Actors age progression. Each line represents the age span for that actor appearing in the movies. ..... 30
**Figure 4.2** : Change in character appearance across the movies. ..... 31
**Figure 4.3** : Precision @ $k$ results for across movies retrieval. This figure is best viewed in color. ..... 36
**Figure 4.4** : Baseline pairwise precision @ $k$ results for across movies retrieval. This figure is best viewed in color. ..... 39
**Figure 4.5** : Pairwise joint metric learning's precision @ $k$ results for across movies retrieval. This figure is best viewed in color..... 41
**Figure 4.6** : Full dimensional Fisher Vectors: Pairwise joint metric learning's precision @ $k$ results for across movies retrieval. This figure is best viewed in color. ..... 42
**Figure 4.7** : Pairwise low-Rank Mahalanobis metric learning's precision @ $k$ results for across movies retrieval. This figure is best viewed in color..... 44
**Figure 4.8** : Pairwise diagonal metric learning's precision @ $k$ results for across movies retrieval. This figure is best viewed in color..... 45
**Figure 4.9** : Confusion matrices of difference between baseline and applied pairwise metric learning mAPs. ..... 46
**Figure 4.10**: FT from all movies: Baseline's precision @ $k$ results for across movies retrieval. This figure is best viewed in color..... 48
**Figure 4.11**: FT from all movies: Joint's precision @ $k$ results for across movies retrieval. This figure is best viewed in color..... 49
**Figure 4.12**: FT from all movies: Joint's precision @ $k$ results for across movies retrieval. This figure is best viewed in color..... 50
**Figure 4.13**: FT from all movies: Low-rank Mahal.'s precision @ $k$ results for across movies retrieval. This figure is best viewed in color..... 52

**Figure 4.14:** FT from all movies: Diagonal metric's precision @ $k$ results for across movies retrieval. This figure is best viewed in color..... 53
**Figure 4.15:** Difference between baseline mAP and applied data from all movies metric learning mAPs. ..... 54
**Figure 4.16:** Sample images for characters that satisfy that SVM training criteria. 55
**Figure 5.1 :** Still images taken in one session under controlled conditions (Studio Settings) [4]. ..... 60
**Figure 5.2 :** Two uncontrolled still images[4]. ..... 60
**Figure 5.3 :** Pipeline of Gabor filters representations for face image [5]. ..... 62
**Figure 5.4 :** Two different aligned faces. ..... 63
**Figure 5.5 :** The three ROC curves plot of the fourth experiment of FRGC using low-rank Mahalanobis metric learning..... 67

**Face Track Retrieval and Recognition Across Age**

## SUMMARY

In this work we focus our study on two subjects: face track retrieval and recognition across age, and the analysis of Fisher vector on face verification. Age invariant face recognition lacks enough researches and dataset based on video. In our study, we present a novel face track dataset called "Harry Potter Movies Aging Dataset (Accio). The dataset is harvested from Harry Potter movie series. It provides face recognition challenges such as variant face pose and expression, illumination, and more importantly a challenge of facial appearance changes due to age progression since the movies span a period of ten years. As a result, this dataset introduces a great environment for studying face track retrieval and recognition across age. Previous datasets conducted on age invariant face recognition use still images (FGNET, MORPH, CACD) either contains small number of data or the data spans small range of years. Our dataset contains large number of face tracks (nearly 38K face tracks) with variant age groups that span an age range between 10 to 88 years. Typically face track recognition is harder than still image face recognition due to effects of tracking, motion blur and pose variation. Each movie of the dataset has different number of face tracks and characters. Nearly 60% of the face tracks are named as they appear in the movies and the rest of face tracks act as distractors or in retrieval evaluation. Our dataset contains great distribution for age group and most of face tracks belong to young characters. This is an important property to study the impact of aging factor since facial appearance changes is more in early ages than older ones. For face track representation, we use the state-of-the-art descriptor: Fisher vector. Fisher vector encoding aggregates large set of local descriptors of all face images into one high dimensional single vector representation.

We define two primary tasks for retrieval and recognition: within and across movies evaluation. (i) *Within movies* task uses the data of the individual movies for training and testing once to evaluate the performance against challenges such as pose and illumination variation. (ii) *Across movies* task adopts evaluation between pairwise movies. Such as training data from HP-1 and testing data from HP-2 or vice versa. The purpose of this task is to assess the impact of aging factor across movies as the years' gap between movies increases. following the definition of evaluation tasks, we suggest three benchmark protocols for the dataset for the evaluation for future studies and researches: (i) *restricted*: that uses only face track query for training and retrieval or recognition models should not use other Accio face tracks or any external data. The aim of this protocol is to compare the performance of face track descriptors. (ii) *Unrestricted*: this protocol allows only for the usage of external data to train learning model for the retrieval and recognition evaluation on Accio dataset. (iii) *Free-for-all* protocol which allows the usage of internal and external data for training.

Using the suggested retrieval and recognition tasks, we introduce an extensive study for face track retrieval and recognition across age. The retrieval evaluation uses two popular performance measures: (ii) Mean Average Precision (mAP) and (ii) precision @k. We apply different experiments on Accio dataset. First Experiment is the across and within movies evaluation on all face tracks with full dimensional Fisher vector. The mAP and Precision @k results of this experiment show clearly the impact of age progression across movies. The performance of within movies evaluation is better than the performance of across movies evaluation due to the facial changes between movies. Following the baseline experiment, and in order to capture the similarity features of face tracks across movies and to improve the performance of this evaluation, we benefit from three metric learning algorithms: (i) *low-rank Mahalanobis metric learning*, (ii) *joint metric learning*, and (iii) *diagonal metric learning*. These metric learning techniques serve two aims: (i) reducing the dimensions of Fisher vectors to make learning more applicable on large datasets, and (ii) increasing the face track retrieval and recognition performance by making the Fisher vectors more discriminative in the new projected subspace.

Prior to the experiments of within and across movies retrieval using metric learning, we reduce the dimensionality of Fisher vectors of face tracks by PCA and then apply the learning algorithms. We use metric learning approaches in two scenarios: (i) metric learning on pairwise movies which combines percentage of face tracks for training a metric model and the uses it for the evaluation on the rest tracks for the two tasks of the evaluation, and (ii) metric learning on data from all movies which combines a percentage of face tracks from all Accio movies and uses them to obtain metric model. Then this model is used to evaluate on the rest of test face tracks for two evaluation tasks. In both scenarios of metric learning, low-rank and joint metric learning improve the results of the evaluation for both measures, mAP and precision @k which reflects their ability to capture variant facial changes through movies and to learn the similarity features between them. However diagonal metric is more basic and leads to slightly worse results than baseline results. For example, the average improvement in between the baseline and the joint learned from all movies is 42.7%. In addition there is a great improvement in precision @k such that the gaps between query and database movies is minimized. This shows the efficiency of these two approaches in learning the similarity between variant facial features across movies. Unlike the baseline where the performance declines for larger k values, results of precision @k is stable for different values in different k values.

The last experiment on Accio dataset is face recognition across age. In this experiment and similar to retrieval tasks, we have within and across movies evaluation. We apply 5 fold across validation test on movies using linear SVMs models. SVM models are obtained from each movies' characters, and we use this model for the validation within the same and across movies. As expected, recognition accuracy in within movies evaluation is higher than accuracy in across movies' evaluation, because of the facial appearance variation between the data of training model and testing data.

In the second part of this work, we evaluate and measure the performance of Fisher vector that we use in face track retrieval and recognition. Fisher vector is successful method and gives great results on face verification task using LFW dataset. However, in our work, the aim is to assess the performance of Fisher vector, using different dataset and features with various parameters. We use FRGC dataset rather than LFW.

Specifically we evaluate its fourth experiment which has one set for training and two sets for test: query and target sets. We use Gabor filter as local descriptors rather than SIFT. We study the effect of spatial, scale and orientation augmentation to the features on the results of face verification. In addition we change the parameters of Fisher vectors such as the number of Gaussian Mixture Models and Feature-PCA dimension. In each experiment, we change one parameter while keeping the others fixed during the evaluation to see the impact of that parameter on the performance of Fisher vectors on face verification task. Since Fisher vector is efficient encoding system in features space, augmentation of the feature information improve the performance greatly. Furthermore, PCA dimension of features has an influence on the results when the dimensions are low, while GMM size does not have big impact on the results.

**Yaşlar Arası Yüz İz Çıkarımı ve Tanıması**

## ÖZET

Videoda yüz tanıma popüler bir araştırma konusudur ve çok yol katedilmiştir. Işıklandırma, çözünürlük ve açı gibi zorluklar çeşitli veri setleri kullanılarak iyice incelenmiştir. Fakat, yüzdeki yaşlanmayı araştırma yönelik video tabanlı hiçbir veri seti bulunmamaktadır. Videolarda yaştan bağımsız yüz tanıma teknikleri geliştirmek ve incelemek için bu zorlu veri setini takdim ediyoruz: "Harry Potter Movies Aging Data Set (Accio)". Veri setimiz sadece açı, aydınlanma ve parazit veriler bakımından zorluklara sahip olmayıp aynı zamanda yüz görünümündeki önemli değişikliklere dair on yıllık bir süreci kapsamaktadır. Yüz çıkarımını bir Harry Potter filminin sadece kendisinden veya diğer Harry Potter filmlerinin herhangi birinden yapmak üzere iki temel görev uyguluyoruz ve harici veri seti kullanımındaki serbestlik açısından değişkenlik gösteren iki farklı protokol kullanılmasını öneriyoruz. Son model bir yüz takibi özniteliği kullanarak çıkarım performansını ana hatlarıyla sunuyoruz. Deneylerimizde yüz çıkarımından alınan sorguyla veritabanı arasındaki yaş farkı arttıkça performansta net bir azalma eğilimi görülmektedir. Videolarda yaştan bağımsız yüz tanıma alanındaki araştırmaların daha ileriye gidebilmesi için veri setimiz kamuya açık olacaktır.

Videolarda yüz tanıma zorlayıcı bir konu olduğu için son yıllarda çok fazla ilgi çekmiştir. TV karakterlerini belirmek, gözetim kamerlarından şüphelileri tanımak gibi birçok pratik kullanımı vardır. Aydınlatma, yüz açısı, ifadesi ve çözünürlük sorunlarını çözmek için önemli gelişmeler kaydedilmiştir. Bu çalışmada, oldukça ihmal edilen yaşlanmaya bağlı yüz görünüm değişiklerinin etkisini incelemeye odaklandık. Yaş varyansyonunu analiz etmek için sabit görüntünün birden fazla veri seti varken (FGNET, MORPH, CACD), bildiğimiz kadarıyla video tabanlı veri seti bulunmamaktadır. Ayrıca, yüz izi kullanarak, yüz takibi nedeniyle haraket bunalıklığı, ve şiddetli poz değişimi nedenleriyle sabit yüz imgesine göre yüz tanıma sorunu daha zor olduğu barizdir. Böyle bir veri seti, yaş değişiminin sorunun video bazlı çözülmesine sağlayacaktır.

Yaştan bağımsız yüz tanıma sistemlerini değerlendirmek ve geliştirmek için geniş ve zorlu yüz izi veri seti sunmaktayız. Veri seti Harry Potter serisinde oluşmaktadır. Filmler on yıllık bir dönemi kapsar ve videoda yüz tanıma diğer zorluklarıyla birlikte yaş varyasyon etkilerini incelemek için güçlü bir ortam sunmaktadır. Ayrıca, veri setinde yaşlanma etkilerini en belirgin olduğu döneme ( genç yüzleri) ait çok sayıda yüz izi bulunmaktadır. Değişik yaş gruplarına sahip olan çok sayıda adlandırılmış karakterlerle birlikte veri seti birçok distraktör olarak adlandırreedığımz ismi olmayan yüz izleri içeriyor.

Bu veri seti on yıllık bir süre içinde yayınlanan sekiz Harry Potter film serisini kullanarak toplanıp düzenlemiştir. Her film kendi içinde birkaç zorluklar içermektedir.

(i) Aydınlatma: birçok parlak ve karanlık sahne vardır. (ii) poz: ön pozlu olmayan yüz izleri ve genelikle aktörler kameraya bakmamaktadır. (iii) çözünürlük: - izler 25'ten 500'a kadar çözünürlüğü olabilir. (iv) distraktörler: yüz izlerin yaklaşık %40'sı arka planda gözüken ve ismi olmayan karakterlere ait ve bu yüz izi tanıma görevinde sorunları yaratabilir. Son olarak, yaşlanmadan dolayı yüzdeki görünüm değişimi sekiz filmi kullanımı iyi bir kaynak olarak görülmektedir.

Yüz izleri elde etmek için ilk başta video çekim sınırları tespit edilmiştir. Her çekim içinde, gibi parçacık filtresi yüz yakalamayla yüz izi elde edilmiştir. Tüm yüz kaydırma ve silindir açılarını kapsayan çoklu poz yüz yakalama yöntemi de kullanılmıştır. Veri seti genç aktörler (yaşı $< 20$ ) için çok sayıda yüz izleri içermektedir. Çoğu yüz görünümü değişiklikleri erken yaşta olduğu için bu önemli bir özelliktir.

Veri setinde yüz izleri temsil etmek için son model olarak bilinen Fisher Vector Faces (FV2) özniteliği kullanılmıştır. Bu makalede yüz tanımak için öznitelik vektörü $(FV^2)$ imgeden yerine yüz izinden çıkartılmış ve buna video pooling denmektedir. Yüz izindeki bütün yüz imgelerinin öznitelik vektörleri bir vektörde birleştirilmiş.

Her filmde ikiye bölünmüş adlandırılmış ve bilinmeyen (distraktör) yüz izlerin sayısı sunulmaktedir. Veri set yaklaşık 38,464 yüz izleri içermektedir. Bunların %59.4 yani 22830 121 farkli karekterlerle filmlerde gözüktüğü gibi adlandırılmış yüz izleridir. Yüzlerin geri kalanı yani %40.6 adlandırılmamış oyunculara ait ve yüz izlerin çıkarımında distraktör olarak kabul edilmiştir. Her Harry Potter film serisinde yüz izleri değişen sayıda bulunmaktadır. Aktörlerin yaşları 10 ile 88 arasındaki yaşları kapsamaktadır. Yaşları filme yayınma tarihine göre hesapladığımız için yaşta hafif bir tutarsızlık olabilir. İlk filmde ve son filmde rol alan aktörlerin yüz izlerini yaş farkı maximum on sene olabilir ve böylece en büyük yaş farkı 10 sene olarak kabul edilmektedir.

İlk olarak filmin aynısını ikinci olarak diğer filmleri kullanarak yüz izi çıkarımı olmak üzere iki temel görev uyguluyoruz:

(i) Filmin kendisi: Bu ayarlarda, her film bireysel olarak görülmüştür ve kendi içinde değerlendirme yapılmıştır. Aynı film içinde adlandırılmış her yüz izi sorgu olarak kullanılırken diğer kalan yüz izleri ( distraktör de dahil) veritabanı olarak oluşturmuştur. Bu ortamda, poz ve aydınlatma gibi tipik yüz tanımı sorunlarına maruz kalan çıkarım performansı değerlendirilmiştir.

(ii) Filmler arasında: By ayarlarda, diğer Harry Potter değişik filmleriyle yüz iz çıkarımını değerlendirerek, aktörler arasında yaş değişininin ektisini analiz edilmektedir. Herhangi film içinden (örneğin HP- 1) her adlandırılmış yüz izi sorgu olarak kullanırken değer bir filmin (örneğin) bütün yüz izleri veritabanı oluşturmaktadır.

LFW ve YouTube Faces setleri gibi verilerin kısıtlı, kısıtsız ve serbest bir şekilde kullanımının ayarlarını öneriyoruz. Görevleri içeren iki protokol öneriyoruz, fakat denetimli ve denetimsiz modelleri öğrenmek için hangi ve ne dereceye verlerin kullanılacağına göre değişir.

(i) Kısıtlı Protokol: Bu protokolda deneylerde harici veriler kullanılmamalıdır. Bunun positif eğitim örneği olarak sadece sorgu izi mevcut olduğuna anlamına gelir. Çıkarım modelleri diğer Accio yüz izleri veya harici iz/imge kullanarak eğitimi kullanılmamalıdır. Bu sadece farklı yüz izin özniteliklerini karşılaştırmak

için veya sorgu genişleme ve alan adaptasyon gibi otomatik ve denetimsiz yöntemleri değerlendirmek için kullanılabilir.

(ii) Kısıtsız Protokol: Bu ortamda, araştırmacılar eğtim modeli elde etmek için harici verilerini kullanmalarına izin verilir. Örneğin, yaşlanma sürecinde yüz görünümü modellemek için harici verileri kullanılabilir. Fakat, aktörlerin modellerini eğitiminde Accio veri setinden herhangi bir yüz izi kullanılmamalıdır.

Son olarak, veriler önceki protokollara göre uymadan farkli bir şekilde kullanılırsa, örneğin yüz izi sınıflandırma amacıyla kullanılması, araştırmacılar buna tüm-serbest düzeni olarak bahsedebilir.

Çıkarım deneyleri iki popüler önlemleri kullanarak değerlendirildi: (i) ortalama kesinlik (Mean Average Precision mAP) (ii) @K Kesinlik, bu de en iyi yüz izlerin sonucunun K değerine karşılık gelir.

Bu çalışma zorlukları, ayarları, ve anahat geliştirilmiş sonuçlarını sunuyoruz. Ilk başta, Kısıtlı protokol kullanarak aynı filme ait yüz iz çıkarımı ( birinci görev) sonuçlarını takdim ediyoruz. Her film için adlandırılmış yüz izleri sorgu olarak kullanırken aynı filmdedki bütün izler veritabnı oluşturmaktadır. Örneğin, ilk film ilk filmde HP-1'de teker teker alınan adlandırılmış 3243 izler ver ve sorgu izi haric bütün 5248 yüz izleri veritabnı olarak kullanılır.

Aynı filme poz ve aydınlatma gibi zorluklarının performansı değerlendirirken, (görev 2) yaşlanma soruna bakıyoruz. Yaş farkı oldukça az oldğunda çıkarım performansını iyi olduğunu varsayarak, aynı zamanda filmler arasında değerlendirme yaptığımız zaman ve yaş farkı arttıkça (orneğin HP-1 vs HP-8) performans kötüleşmesini görüyörüz.

Sonuçlar bölümünde, sonuçların ana hattını geliştirmek amacıyla ve görev 1 ve 2 sonuçlarınında farkını azaltmak için metric öğrenme tekniklerinden yararlanyoruz. Deneylerde elde edilen yüksek boyutlu FV metric öğrenme kullanarak sıkıştırılıp ayrıt edici bir temsil ortaya çıkarmaktadır. Sıkıştırma doğrusal projeksiyon kullanılarak gerçekleştirilir ve iki amaca hizmet eder:
(i) Yüz öznitelikleri azaltır, büyük ölçüde ve büyük ölçekli veri setleri için geçerli hale gelir.
(ii) Alt uzay üzerine projeksiyon, ayrımcı Öklid mesafe ile tanıma performansını artırır Bu çalışmada da Mahalanobis metrik yanında da diyagonal metrik ve joint metric kullanılmıştır.

Metric öğrenme uygulamadan önce de, FV boyutlarını azatlmak için üzerinde PCA uyguluyoruz. Metric öğrenme teknikleri iki farklı şekilde kullandık:

(i) Çift film üzerinde metrik öğrenme: iki filmden eğitim için bir oran veri alınıyor ve daha sonra bunlarla metrik model elde ediliyor. Bu model kullanarak, kalan veri üzerinde test ediyoruz.

(ii) Bütün filmler üzerinde metrik öğrenme: bu tip kullanmada, bütün filmlerin yüz izlerinde bir oran (örnek: 10%) toplanıp ve üzerinde metric öğrenme eğitimi yapıyoruz. Kalan yüz izleri iki çıkarım görevlerini uyguluyoruz.

İki senaryoda, joint ve low-rank Mahalanobis teknikleri, iki çıkarım görevlerinde mAP ve kesinlik sonuçlarını artırmaktadır. Bu iki öğrenme metodlarının sonuçların arasındaki farkı azalmakta ve değerlerini oldukça geliştirmektedir. Bunun nedeni,

metric öğrenme yüzdeki yaştan dolayı bigisini alıp özniteliklerini aynı domaina sıkıştırıp benzerliklerini öğrenmesinden kaynaklanmaktadır. Fakat diyagonal metrik öğrenme, öncekilerine göre daha basit olduğu için, yaştan dolayı özniteliklerinin farkını öğrenmesi için yeterli değildir. Bu yüzden, bu teknik anahat sonuçlarından daha kötü olduğunu görüyörüz.

Ayrıca linear SVM kullanarak yüz iz tanıma deneyleri gerçekleştirdik. Yaş etkisi görmek ve performansı değerlendirmek için iki çıkarım görevlerini de tanıma için uyguluyoruz. Değerlendirme için 5 çapraz doğrulama (5 fold cross valildation) yöntemi kullanılmaktadır. İlk başta, değerlendirme yapılmadan önce, en az 3 filmlerde ve 50 yüz izi olan karakterler değerlendirmelerde dâhil edilmiştir. Kullanılan karakterler için mutli-class SVM eğitimi ve değerlendirilmesi uygulandı. Örneğin, ilk Harry Potter filminde 23 karakter için SVM eğitimi yapıldı ve daha sonra test için elde edilen SVM modelleri hem aynı filmde hem de diğer filmlerde test edilmektedir. Test sonuçlarına göre eğitim seti ile test seti aynı filmden seçildiğinde başarım yüksek çıkmıştır. Test verisi eğitilen modelden farklı bir filmden seçildiğinde ise başarımın düştüğü görülmüştür. Bu da veri setimizde yaş faktörü ne kadar güçlü olduğunu göstermektedir.

Tezin ikinci çalışmasında, kullandığımız ve Fisher vektör yönteminin başarmını ölçmek için analiz ediyoruz. Bu yöntemi kullanarak LFW kısıtlı deney setinde en iyi sonuçlardan birini vermektedir. Kısıtlı olmayan deneylerde de oldukça başaralı ve yüksek sonuçlar vermektedir. "Fisher Vector Faces in the Wild" makalesinde önerilen yöntemi SIFT öznitelik ve metrik öğrenme yöntemleri kullanarak farklı öznitelikleri, ve parameterleri test edilmektdedir.. Önerilen FV yöntemi, farklı veri seti üzerinde nasıl çalıştığını ve performans nasıl etkilendiğini ölçmek için, bu yöntemi Face Recognition Grand Challange (FRGG) veri setinin dördüncü deney üzerinde değerlendirdik. FRGC veri setinde eğitim kümesi 222 kişiye ait 12,776 imgeden oluşur. Test için galerisi ve prob setleri kişi başına tek hareketsiz görüntülerden oluşur. Bu deneyde test için 466 kişiye ait, sorgu veri setinde kontrolsüz 8014 ve hedef veri setinde 16028 kontrolü imgeden oluşur. Genelde performans, ROC eğrisinde 0,1%'inde doğrulma oranı (FAR) olarak elde edilir. Üç tane ROC eğrisi var, ROC 1 imgeler aynı dönemde, ROC 2 imgeler aynı senede ve ROC 3 imgeler iki dönem arasında çekildiğini ifade ediliyor. Sonuçlar, ROC Eğrisinin altında kalan alan (Area Under the Curve) olarak değerlendirildi.

FV yüz tanıma için SIFT'ten farklı öznitelik kullanıldığında performansı görmek için Gabor filter öznitelikleri kullandık. Her bloktan 256 boyutlu öznitelik vektörüne, Gabor filtresine ait hangi ölçek, oryantasyon ve bloğun x ve y bilgileri eklendiği halde, performans nasıl değiştiğini incelendi. Özniteliklerin bilgileri eklendiğinde iyi sonucu FV vermektedir. Bu bilgiler eklenmediği halde sistemin başarımı düşük olduğunu görüyürüz.

Yüz imgesinin her bloğundan, 256 boyutlu öznitelik vektörü elde edildikten sonra, PCA tekniğin kullanarak daha düşük boyutlu vektör haline dönüştürüldü. Bu deneyde, sadece PCA boyutu değiştirirken, diğer deney parametreleri GMM, Örtüşme, Eklenme sabit tutuldu. Sistemin performansı düşük boyutlarda daha iyi çalıştığını fark edilmektedir. Yapılan önceki deneylerden sonra, en iyi sonuç veren parametreleri seçip bu parametreleri kullandıktan sonra, sistemin başarımı karekök normalleştirme, örtüşmeyen bloklar, az PCA boyut ve ölçek, oryantasyon ve mekânsal bilgileri eklenme parametreleriyle en iyi sonucu verdiğini görülmektedir.

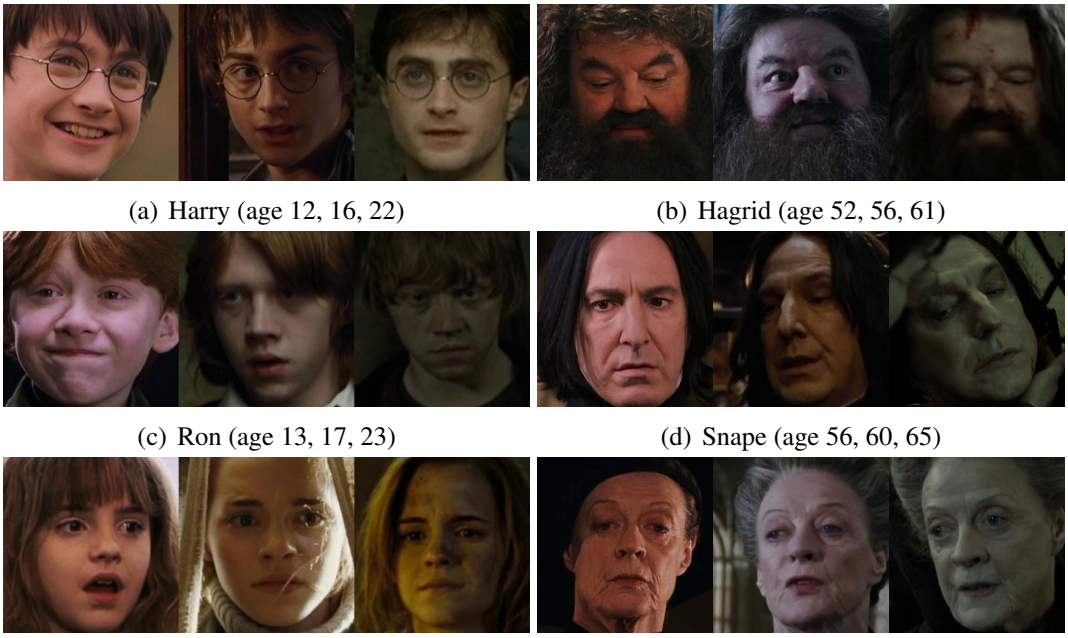
# 1. INTRODUCTION

Face track retrieval and recognition in videos has received lots of attention in recent years, and is a hard subject in computer vision [6–8]. It has many practical applications such as security surveillance [9], finding missing children and recognizing people across age in videos s [7]. However it still lacks enough interest and focus. Technical challenges such as illumination, occlusion and variant pose and face expression have been studied well in the past few decades. Nonetheless, face recognition still a hard problem under real world conditions. In addition, challenge such as facial appearance changes due to aging is not well analyzed and solved. In this thesis we study and analyze face verification and face track retrieval and recognition across age using state-of-art Fisher vector representation.

In the first part, we study and analysis face track retrieval and recognition across age and present challenging video based dataset using Harry Potter movie series called "Harry Potter Movies Aging Dataset (Accio[1])". Due to the lack of video based face recognition dataset with age variation, we introduce and make available a large, and challenging this face track dataset ( Accio: see Sec. 4.1). The aim of this work is to investigate the facial appearance challenge caused by aging. This will help to further evaluate and develop age invariant face recognition systems, specifically in videos. To the best of our knowledge, the dataset is gathered from the eight movies of the Harry Potter series.

Harry potter movies dataset contains many challenges such as illumination, occlusion, video quality and other face recognition challenges. Since the movies span a period of ten years, dataset present strong settings to study the effects of age variation along with all other challenges of video face recognition. As a result it provides an interesting and excellent case study to improve and get face recognition techniques that would overcome challenges due to the time gap between videos taken in different years for some identities. The set includes diverse identities that appeared in Harry Potter

[1]The Summoning Charm *Accio* in the Harry Potter series retrieves an object at a distance, here used to signify face track retrieval.



(a) Harry (age 12, 16, 22) (b) Hagrid (age 52, 56, 61)

(c) Ron (age 13, 17, 23) (d) Snape (age 56, 60, 65)

(e) Hermione (age 11, 15, 21) (f) McGonagall (age 67, 71, 76)

**Figure 1.1**: Change in character appearance across the movies.

movies and shows distinct facial changes through eight movies. It shifts the scope of previous still image based sets (FGNET [10], MORPH [11], CACD [12]) to a broad and face track age invariant face recognition due to its size and its source. Often, face track recognition proves to be harder than still image face recognition due to effects of tracking, motion blur and pose variation. Such a dataset will help untangle the challenge of age variation catering specially to the broadcast video domain. Furthermore, the dataset features a large number of face tracks with young faces (age $< 20$) where the effects of aging on facial appearance are most prominent. Along with a large number of named characters and varying age groups, the dataset also contains many unknown tracks which act as distractors. Fig. 5.2 shows the changes in facial appearance for six characters taken from movies first, fourth and eighth Harry Potter movies. Note that these images are deliberately chosen to have frontal pose and decent illumination to highlight the variation in their age.

As a benchmark for the face track retrieval and recognition across age, we evaluate the following scenarios:

- Face track retrieval within movies
- Face track retrieval across movies

- Face track recognition

Similar to LFW [13] or YouTube Faces [14] we also propose the usage of data in a restricted, unrestricted setting, and free-for-all protocol that allow using the dataset's face tracks for learning goals. Sec. 4.2 presents more details about the challenges, and baseline and improved results benefiting three metric learning:

- Low-rank Joint Metric Learning[15, 16]
- Low-rank Mahalanobis Metric Learning [17]
- Diagonal Metric Learning

To improve the baseline results, we use metric learning with two different settings: (i) *metric learning on pairwise movies:* which obtains a metric learning model from Accio dataset's two movies, for example first and second movies and test on the rest of tracks. (ii) *Metric learning from all movies:* which gets a metric learning model using data from all Accio dataset's movies and test on the rest of tracks. We notice an improvement over the baseline results using both metric learning scenarios and with the two metric learning methods: low-rank and joint metric learning.

For face track recognition across movies we use Linear Support Vector Machine (SVMs) as a classifier and we present the recognition accuracies of within and across movies recognition. In addition, we present recognition results of specific and major characters that appear in all movies and show great facial changes across movies due to aging.

To promote researchers on face track media retrieval we include a baseline retrieval results and we will make the dataset including details and experiment settings available to download.

In the second part of this thesis we analyze the performance of Fisher vectors by evaluating the fourth experiment of the Face Recognition Grand Challenge (FRGC) dataset. The aim is to see how Fisher vectors for face verification would perform using different dataset, features and settings. In this work, we use the pipeline of Fisher vectors for face verification as suggested and implemented in [3] which uses root Dense-SIFT as local descriptors, Gaussian Mixture Models (GMM) as visual words

and then encode these local features by aggregating them into one single Fisher vector. Following this we use Gabor filters as local descriptors for face images as proposed in [5] prior the steps of GMM training and Fisher vectors encoding. In addition, to study the performance of Fisher vectors related to specific settings, we use variant face alignments, parameters of features, Principal Component Analysis(PCA) dimensions, and GMM sizes.

### 1.1 Motivation and Hypothesis

Our motivation behind the studying of age invariant face track retrieval and recognition in videos is due to the lack of work in this area and a video based dataset that deal with the facial changes due to aging factor. The existing studies use still images not videos and they either contains small number of data or the data spans small range of years. Our dataset introduces challenging settings primarily based on age challenge since the source videos (Harry Potter movie series) span ten years. Additional challenges such as large illumination changes, occlusion, face pose and expression also exist. Furthermore, we want to analyze the state-of-art method Fisher vectors on this task combined with successful metric learning algorithms.

Prior to learning a model on Accio face tracks, we expect within movies performance to be always high compared to the across movies results. Cross movies results are obtained from the retrieval of pair movies, for example first and second movies. Figure 1.2 illustrates the dominant matrix of within and across movies evaluation results. In the figure the diagonal elements are dominant with highest energy since they present within movies results where retrieval and classification is subjected only to standard challenges of face recognition in videos such as illumination and pose variation. The off diagonal elements are less dominant and has less energy, since they provide the across movies results where face track retrieval and recognition are subjected to the aging factor across age and to the other challenges of face recognition in videos. In our experiments, this hypothesis is proved and mostly the performance is high in the diagonal elements and around it, and it decreases as we go farther from the diagonal elements where the years gap increases due to the aforementioned reasons.

Face verification is a popular task that received considerable attention during recent years. Many studies achieved good performance using datasets such as LFW. Recently

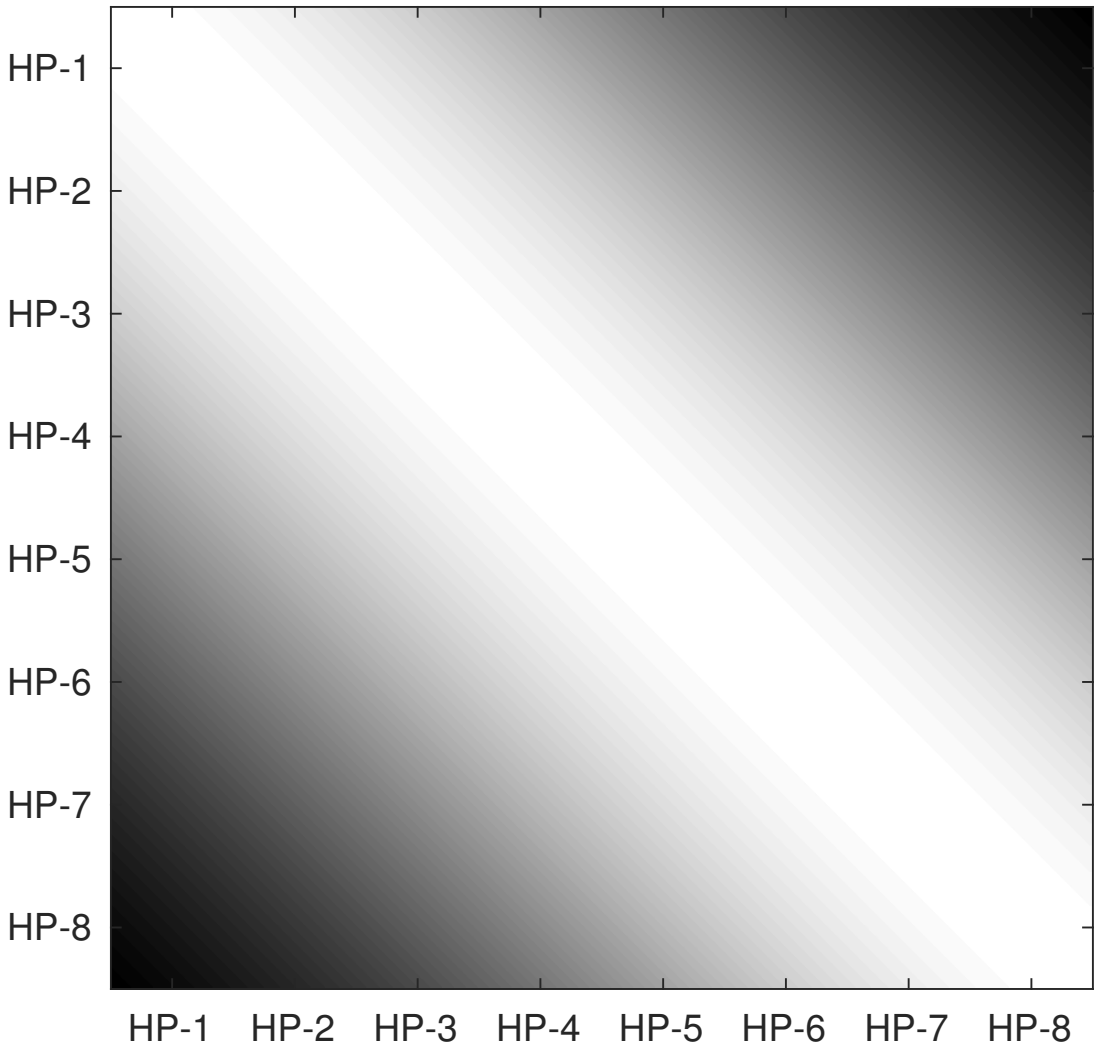



**Figure 1.2**: Across movies evaluation hypothesis: Dominant matrix.

LFW is the most popular evaluation benchmark and widely used to test the proposed and published methods for face verification. Fisher Vector Faces in the Wild [3] is one of the successful approaches which use Fisher vector encoding for large sets of local descriptors. It achieves good results under LFW's restricted and unrestricted settings. Nonetheless, face verification is still considered a hard task under real world conditions. Our motivation of the Fisher vector analysis in this work is to evaluate the performance of this well-known technique and use it during our study of face track retrieval and recognition across age in videos. The aim is to measure its performance using different datasets other than LFW, and different local descriptors such as Gabor filters with varying settings. This work presents a detailed analysis given in chapter 5.

## 1.2 Problem Definition

In face track retrieval and recognition, there are two tasks: (i) *within movies* and (ii) *across movies* retrieval and recognition. For example, In within movies retrieval, named tracks in each movie are used as query once, while the rest are used as database. In face recognition within movies, supervised model such as linear SVM is trained on part of face tracks and used for test on the rest part of face tracks. In this task, a regular track retrieval and recognition are applied which regards standard technical challenges such as illumination and variant face pose. In across movies retrieval and recognition task, it is applied on pairwise movies for example named tracks queries are taken from the first movie and the database from the second movie's all tracks. Similarly, for across movies face track recognition, a model obtained from a movie is used for test on another movie. In this task retrieval and recognition evaluations measure the impact of aging progression across movies. It is important to mention that, in face track retrieval for the two tasks, only we select only named tracks as queries while for the database we use all tracks including the distractors. In face track recognition, only named characters which have at least 50 face tracks and appear in at least three movies are included in the evaluation.

In face recognition, face verification is the task where to decide whether a given pair images are for the same person or not. The aim is to maximize the distance between negative pair images of different people and minimize the distance between positive pair images of the same person.

## 1.3 Thesis Outline

Chapter 1 introduces the two tasks of this work: (i) *face track retrieval and recognition across age*, and (ii) *analysis of Fisher vectors on face verification*. It gives an explanation, the motivation, and the definition of the two tasks in this work.

Chapter 2 presents the previous work on both face verification and age invariant face recognition. In this chapter we give a brief summary of previous studies on face verification. In addition we outline related work on age invariant face recognition and

explain previously used datasets for age invariant face recognition. We also illustrate some work on face track retrieval.

Chapter 3 deals with the theoretical foundations related to face identification on still images and videos. First section presents shot boundary detection which simplify the process of face detection and tracking in videos. Then we explain the pipeline of face identification which starts with face detection and tracking, then face alignment, and finally feature extraction. For these tasks, we summarize the approaches we use in this work such as Modified Census Transformation (MCT) for face detection, particle filters for face tracking, similarity transformation using eye centers for face alignment, and Fisher vectors for feature extraction. Following the pipeline of face identification, this chapter describes the learning algorithms used in this work: *(i) joint metric learning*, *(ii) low-rank Mahalanobis metric learning*, *(iii)diagonal metric learning*, and *(iv) linear support vector machines (SVMs)*. Finally we discuss the evaluation metrics that we apply to measure the performance of face track retrieval, recognition and verification tasks.

Chapter 4 explains the first part of this work: face track retrieval and recognition across age using Accio dataset. It starts by providing all the details of the novel dataset such as the number of face tracks, characters, age range and some statistics on age groups. In addition, in this chapter we compare the dataset with previous datasets for age invariant face recognition based on still images. Then we describe Fisher vector method which is used for face track description. The chapter also introduces across and within movies evaluation tasks on the dataset with restricted and unrestricted settings. Then we describe the experiments using different metric learning with two variant scenarios: metric learning on pairwise movies which combines the two movies training face tracks to obtain a metric learning model, and metric learning based on training face tracks from all movies to produce metric learning model. Then we provide all experiments' results and compares them to the baseline evaluation. Last section of this chapter presents the baseline results of face track recognition by applying linear SMV learning. Finally we give the recognition F1 score of some major characters from the movies.

Finally chapter 5 gives the analysis of Fisher vector in face verification using FRGC dataset, SIFT features, and Gabor filters. It starts by explaining FRGC dataset and its

fourth experiment. Then we describe Fisher vector with SIFT features and provides the evaluation results. As a part of Fisher vector analysis, we adopt Gabor filters as local descriptors for Fisher vector and shows the effects of using these features with various Fisher vector's parameters on face verification task. The chapter provides all details of the experiments and the comparison of different scenarios.

Chapter 6 summarizes the contribution of this thesis and presents a direction for potential future research related to the subject.

# 2. RELATED WORK

## 2.1 Face Identification

Face identification typically includes the following pipeline steps: face detection, facial landmark detection, face alignment, face description, and finally face classification. Nonetheless, studies usually focus separately on each of these steps. Viola Jones face detector[18] is widely used since it is open source and has implementation in OpenCV library[19]. However in this thesis we adopt Modified Census Transformation (MCT) [20] for face detection. These two methods are early and widely used for face detection and they are both based on scanning window and cascade classifiers. In facial landmark detection, Supervised Descent Method [21] is successful face shape regression technique. It begins with an initial $S_0$ face shape and progressively predicts the final face shape in iterative way. For our study we benefit from facial landmark provided by SDM mainly for face alignment.

Various methods are used for face alignment. Similarity transformation based on facial points aligns faces such that they are fit in a similar size and position. For example [22] applies similarity transformation using 9 facial points to transform a face to a canonical frame.

There are many studies for face description. The aim is to find robust and discriminative face representation such that faces of negative pairs of different people are more separated and positive face pairs of same person are closer in feature space. These studies include [3, 5, 23, 24].

For face classification, Support Vector Machines (SVMs) are widely used [25–27]. In addition many metric learning algorithms have been proposed recently [15, 28]. In discriminative metric learning there are two major aims: (i) reducing the dimensionality of face descriptors and (ii) preserving the discriminative features. This is done by learning a low-rank linear projection matrix $W$ on the descriptors. The metric is learned from given positive and negative pair images which captures the

similarity and the difference between the pair images such that the Euclidean distance between positive pair images is minimized while it is maximized when the pair images belong to different people. For example learning a Mahalanobis matrix $M = W^T W$ which is a problem with convex formulation [17]. Learning $M$ is practical only with low dimensional feature vectors. As a results, usually a dimensionality reduction, for example by PCA, is used to reduce dimensions into applicable numbers and then performing metric learning.

Previously there are several works for face verification, our work is guided by Fisher vector faces in the wild [3] and multi-resolution local appearance-based face verification [5] papers. Our case study of face verification is an analysis of the performance of fisher vector using different dataset and features. In this thesis we use Gabor and SIFT features for training and encoding Fisher Vectors. Then we evaluate the method on the fourth experiment of FRGC dataset using variant parameters.

### 2.2 Data Sets

There are several datasets based on still images to study face verification problem such as LFW [13], FRGC [4] and MORHP [11]. You-Tube Faces [14] is also popular dataset for studying face recognition in video. Nonetheless, these datasets do not consider the face pairs matching cross ages. Prior face recognition datasets which deal with the problem of aging are all using still images [10–12, 29]. Datasets for video face track recognition are usually collected from TV series or one movie [6, 7] which means they do not have age variation. To best of our knowledge, in this thesis we present the first video based face track retrieval and recognition dataset which is gathered from Harry Potter movie series and has challenging settings. More importantly it shows clear trends of effects of the aging factor.

### 2.3 Face Track Retrieval

Among the first researches of face recognition on videos, most works were limited to the frontal faces [30] and few of them included profile faces [31–34]. Apart from face poses, representations of face tracks generally followed two approaches, the first one represents face in a track individually and uses face track as a set of images. The benefit of these methods is to allow using the developed representations of still images

for face tracks. The second type of approaches represents a face track as a whole to a single feature vector such as [8], 3D model or manifold. These techniques can be generative or statistical. For example, [35, 36] try to build 3D model from the data and use it to explain the new data. Statistical methods such as [30, 37], learn local features across the face track images and then represent them into BoW model or GMM model. This leads to a single feature vector representation of the whole images of a face track.

As an example of early face track retrieval works is [30], where they detect frontal faces, and for tracking they use region-tracking technique and then adopt SIFT as local descriptor around 4 facial landmarks for verification. This helps to find corresponding shots in a video where a person of query track appears. Similarly, [38] includes non-frontal faces and [39] provides a compact descriptor for face track matching. Everingham et al. [7] paper made an evolutionary step and since then most researches are about on the assigning a name to all face tracks (for example [6]). This method requires finding transcripts which is not always possible.

### 2.4 Age Invariant Face Recognition

There are multiple studies which consider the problem of aging in face recognition. While age estimation [40–46] and simulation [47–53]are extensively studied, researches that explicitly deal with developing age invariant face recognition systems are not enough. Researches which consider the problem of aging in face recognition can be roughly divided into two subcategories: Generative and discriminative techniques. (i) *Generative* approaches [50, 54–56] resemble the face image at a certain age and then perform face recognition. Usually this is done by building 2-D or 3-D generative model to simulate the aging process for face matching. For example[56] uses facial landmarks to estimate the facial growth parameters and predicts the appearance at certain age. However these approaches have drawbacks in many aspects, they are based on parametric assumptions that may results in unrealistic simulation and they have high complexity during computation.

(ii) *Discriminative* approaches: [12, 57–59] use learning methods to build age invariant face recognition systems. For example, Chen et al. [12] uses a reference set representations for each person, his/her age, and for specific facial landmarks. They

model the task as least square and pool across the age space to obtain an age-invariant representation for face recognition.

Different from previous approaches, some studies of modeling the facial appearance through latent space. For example,Gong et al. [60] proposes to use two latent factors, an identity factor for age-invariant face recognition, and an age factor affected by aging to capture the aging process features. Then observed facial appearance is modeled by the fusion of the two factors, and separating the observation enables performing recognition on the identity part.

# 3. THEORETICAL FOUNDATION

The aim of the thesis is to address challenging face track retrieval and recognition across ages using Harry Potter Movies and also study and analyze the efficiency of Fisher vectors using different features and datasets with variant settings. In this chapter, we explain theoretical background, machine learning and computer vision technologies used in the thesis. We describe the techniques adopted to obtain Accio dataset such as shot boundary detection, face detection and tracking, the descriptor used for representation of still images and face tracks. Finally we discuss the learning algorithms and evaluation methods of the tasks.

## 3.1 Shot Boundary Detection

In order to simplify the process of face recognition, face detection and face tracking in videos, shots of a video are detected and faces on every shot are tracked to obtain face tracks. Shot represents a sequence of frames taken by a camera for a continuous action. Multiple of logical and sequence shots construct what is called scene in a video [61]. Shot boundary is the transition between two shots that can be hard cuts or soft (fade, dissolve, wipe ...etc.). Figure 3.1 illustrates two shots of one scene.

**Figure 3.1**: Shot boundary example [1].

There are many approaches to detect shot boundaries. In this work, we use Displaced Frame Difference (DFD) [62] for shot boundary detection since it achieves an acceptable accuracy with a good speed. DFD works well in case when we the motion between two consecutive frames is not too much. For a sequence of $N$ frames $F(r,1)....F(r,t)...F(r,N)$, where $r$ indicates the pixel position, the comparison of two

consecutive frames in one shot can be given by the following sum formula

$$FD(t+1) = \|F(r,t+1) - F(r,t)\| \tag{3.1}$$

It can be decided whether the two frames belong to the same shot or not by using a learned threshold of cut locations. If $FD(t+1)$ is above a certain threshold or sufficiently high, it indicates that the frame is different from the previous one a time $t$ and a shot boundary is accepted.

### 3.2 Face Detection and Tracking

Face detection is essential step in face recognition pipeline. It is the first step in the face recognition systems. Face detection is extensively studied and searched during the past few decades. There have been many methods proposed for face detection. Viola Jones [18] had made an evolutionary step in face detection. It uses Haar features selection, creating integral image, Adaboost training and cascade classifiers. It works well on detecting frontal faces.

In Accio dataset, to obtain face tracks for all movies, we use MCT for face detection [2] and we follow [6] for face tracking which uses particle filter based on tracking by detection. This enables obtaining face tracks of each character within a shot. MCT is robust against illumination and pose variations with very high rate of detection rate. It is a practical method and efficiently fast.

#### 3.2.1 Modified census transformation

Census Transform(CT) is a structure kernel feature to capture the local spatial structure of the image. It is basically computed for each pixel by comparing a pixel with its neighbors of a kernel size $3 \times 3$. Let $N(x)$ be spatial neighborhood of a central pixel $x$ which we want to compute its CT feature. The obtained CT feature is a bit of binary string resulted from the logical function computed as,

$$\lambda(I(x),I(y)) = \begin{cases} 1 & \text{if } I(x) < I(y) \\ 0 & \text{if } I(x) \geqslant I(y) \end{cases} \tag{3.2}$$

This function checks the intensity of the neighbor pixel, if it is less than $x$ then it gives 1, and 0 otherwise. The CT transformation is obtained by

$$CT(x) = \oplus_{y \in N(x)} \lambda(I(x),I(y)) \tag{3.3}$$

Where $\oplus$ refers to the concatenation operation for the logical function to generate CT string. Since there is 8 neighbors, the CT feature has a range between $0-255$ values.

MCT is a modified version of the aforementioned CT. The previous method results in 256 subsets of the structure kernel of $3 \times 3$spatial neighborhood, however in order to compute all structure kernels, the central pixel should be included in the logical and comparison function rather than fixing its value to zero.

### 3.2.2 Boosting and cascade classifiers

In face detection the MCT detector analyzes patches $W$ of $18 \times 18$ size. During boosting and cascade classifiers steps, each window is decided to be a face or a background. The decision is made in a sequence of tests for efficiency in boosting process which combines several classifiers. The window can be rejected after any test stage. In this thesis, the system consist of four stages that can be displayed in 3.2. As shown in the figure, in first face, block white dots show the positions of the elementary classifiers in the selected window. These white dots are increased in each stage. Early three stages can classify the given window as a background, however only the final stage can decide whether the given window is a face or not.

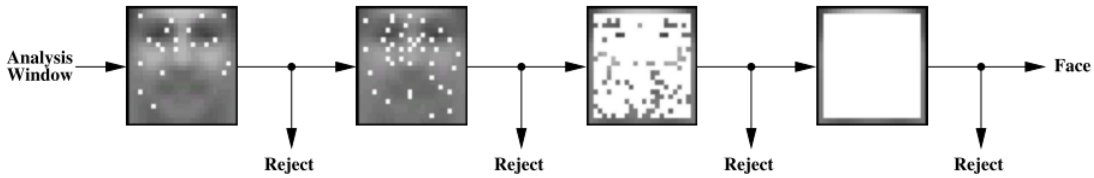



**Figure 3.2**: Face detection with MCT features and cascade classifiers [2].

### 3.2.3 Face tracking

Face tracking is essential since face detection in every sequence of frames in a video is not efficient and it degrades the performance of the system. Tracking faces' motion in a video sequence that have variant expressions, scales and illumination is a hard task. Face tracking has many usages such as human surveillance, interaction application, and face or object model structuring. Many well-known algorithms have been developed in order to track objects for examples mean shift, Kalman filter, and particle filtering. Face detection, tracking and labeling tool provided in CVHCI laboratory is used to

obtain face tracks in this work. Following [6] particle filter based on face detection is adopted in this tool. Particle filtering is a recursive Bayesian filter. It is used to estimate the posterior density of the state space $X_t$ given the observation variables $Z_t$, $p(X_t|Z_t)$. Particle filter is based on using particles to represent the posterior density of the state space. It is well known and efficient for nonlinear and non-Gaussian estimations[63]. In the face tracking and annotating tool provided by CVHCI laboratory, MCT face detector is integrated with particle filter based on detection. At first, faces are detected then the locations around detected faces are evaluated since they are likely to contain the faces. Each tracked face by particle filter has particles with state $x = \{x.y, s, \alpha\}$, where $(x, y)$ gives particle location in the 2-D image, $s$ is its size and $\alpha$ is the yaw angle of the particle.

Following shot boundary detections in videos, faces within a shot are detected in first frame by scanning the whole frame at varying scale and head pose angles. These detections are initialized for face tracks, if an occlusion is emerged due to the intervention of two face tracks, one of the tracks is terminated. MCT is run to evaluate and assign scores for the generated particles. The final decision of the tracked face position is given based on the weighted average of the individual particles. A face track is terminated when no faces are found within certain number of detections. The initialization of the face tracks are performed every 5 frames to detect new faces that may appear.

Figure 3.3 shows an examples of face tracks that contain different number of images. These tracks are obtained from Harry potter movies which are the source dataset (Accio) for face track retrieval and recognition tasks in this work.

### 3.2.4 Face alignment

Face alignment is important step in face recognition. It is the process of aligning facial landmarks, for instance: eyes, nose, mouth, and chin to canonical frame [64]. This process fixes the landmarks positions in aligned images and it is carried out by similarity transformation. In our work we use facial landmarks provided by SDM landmark detector and perform similarity transformation that aligns based on eye centers coordinates. Examples of aligned faces according to the distance between eyes and row distances are shown in figure 3.4

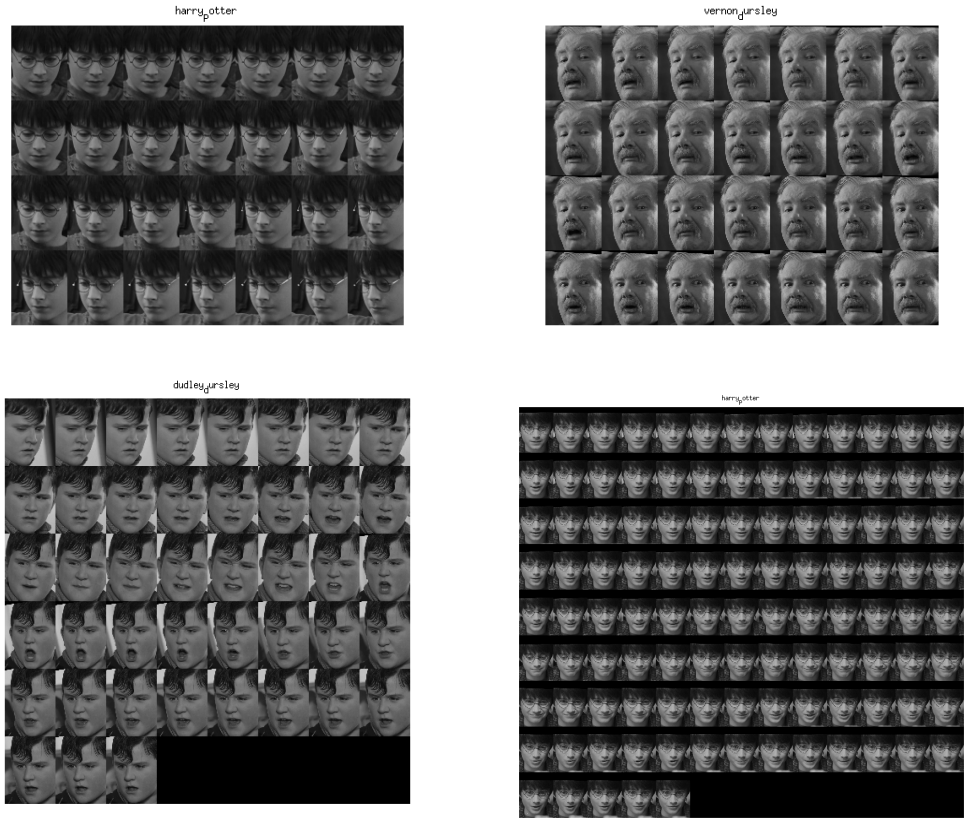



**Figure 3.3**: Face tracks obtained by face tracker in Harry Potter movie series.

### 3.3 Face Image and Track Representation

This work adopts Fisher vector implementation in [3, 8] in order to obtain the face images and face tracks descriptors. For Fisher Vectors analysis, a apart from the Fisher vector pipeline suggested in th paper, GABOR features were used in order to test and evaluate the performance of Fisher Vectors. This section explain Fisher vector implementation for face recognition. The pipeline typically starts by obtaining face images and tracks as explained in the previous sections, then extracting the dense root-SIFT features, dimensionality reduction of the SIFT features by PCA, augmenting of spatial information of SIFT blocks, training Gaussian Mixture Models, and then encoding the large sets of SIFT features by aggregating them into a single Fisher vector using GMM. We illustrate the aforementioned steps of fisher vector implementation in figure 3.5

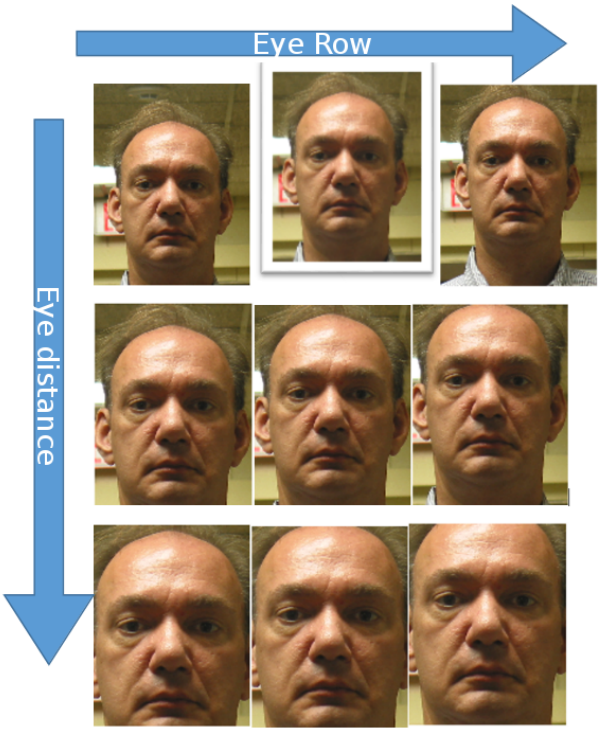



**Figure 3.4**: Aligned faces with different eye distances and eye rows.

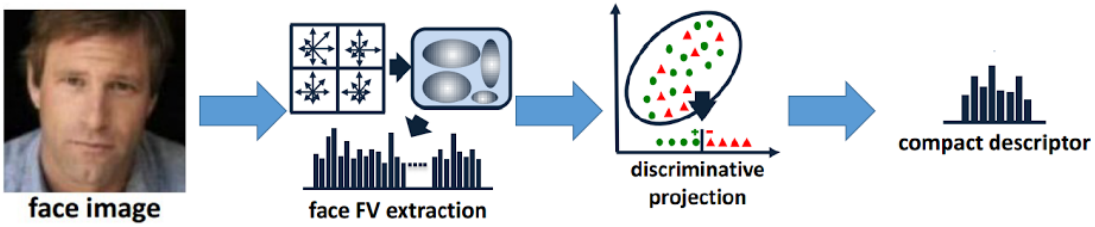



**Figure 3.5**: Fisher vector implementation pipeline [3].

### 3.3.1 Fisher vector for face recognition

**Features:** Fisher vector implementation starts by extracting dense features from face image patches. We use two features in this work: (i) *SIFT features* and (ii) *Gabor filters*.

**Scale Invariant Features Transformation:** Instead of sampling features sparsely around certain positions detected by facial landmarks detector, [3] implementation suggested extracting features on the face densely in scale and space. Root-SIFT features are computed on patches of $24 \times 24$ pixel resolution sampled with step of size 1. Computation is repeated on five scales with scale factor equal to $\sqrt{2}$. This computation leads to a set of $3887 \times 128$- dimensional descriptor per face's image

with a size equal to $80 \times 60$. Prior to the Fisher vector encoding of features, descriptor dimensions are reduced by PCA from 128 to 64 to make the extracted features amenable to Fisher vector encoding and applicable for learning on large sets.

**Gabor Filters:** In addition to SIFT features, in this work we use Gabor filters as local descriptor for face representation preceding to Fisher vector encoding. 2-D Gabor wavelet are one of the successful local descriptors for face representation since it has biological relevance to human visual system. They decompose the input images into several scales and orientations. This representation is widely used in computer vision due to its efficient localization feature in both frequency and spatial domain.

Gabor filters are bandpass filters which are a two dimensional Gaussian kernel function modulated by sinusoidal planar wave. This complex filter is defined as following:

$$\Psi_{u,v} = \frac{\|k_{u,v}\|}{\sigma^2} e^{(-\|k_{u,v}\|^2 \|z^2\|/2\sigma^2)} [e^{i\overrightarrow{k_{u,v}z}} - e^{-\sigma^2/2}] \tag{3.4}$$

where $k_{u,v} = k_v e^{i\phi_u}$ , $k_v = \frac{k_m ax}{f^v}$ are the frequency parameter and $\Phi - \frac{u\pi}{8}, \Phi \in [0, \pi]$ is the orientations parameter. The defined sinusoid wave in 3.4 is activated by frequency information of the image. This leads to the capturing of frequency information near the frequency of the sinusoidal wave and ignoring the other frequency information. The previous frequency representation is due to the Gaussian envelope that ensure the convolution is dominant in image region that is close to the Gabor response [5].

Gabor filters are designed by multiple of orientations and scales which can act as number of dilations and rotations by using variant u, and v parameters for example in a range of [0 ,7] and [0 ,4] respectively. As a result we obtain 40 Gabor wavelets with 8 orientations and 5 scales. The Gabor wavelets can be shown in figure 3.6. Each wavelet deals with specific edge information for face representation. Usually a filter bank consisting of Gabor filters with different scales and orientations is created rather than computing the wavelets each time. Then the filters are convolved with the input image resulting in what is called Gabor-space.

After filtering we obtain 40 Gabor magnitude images (GMIs) that correspond to different scales and orientations. These images are divided into $N$ non-overlapped blocks, each belongs to local patch of a face image.

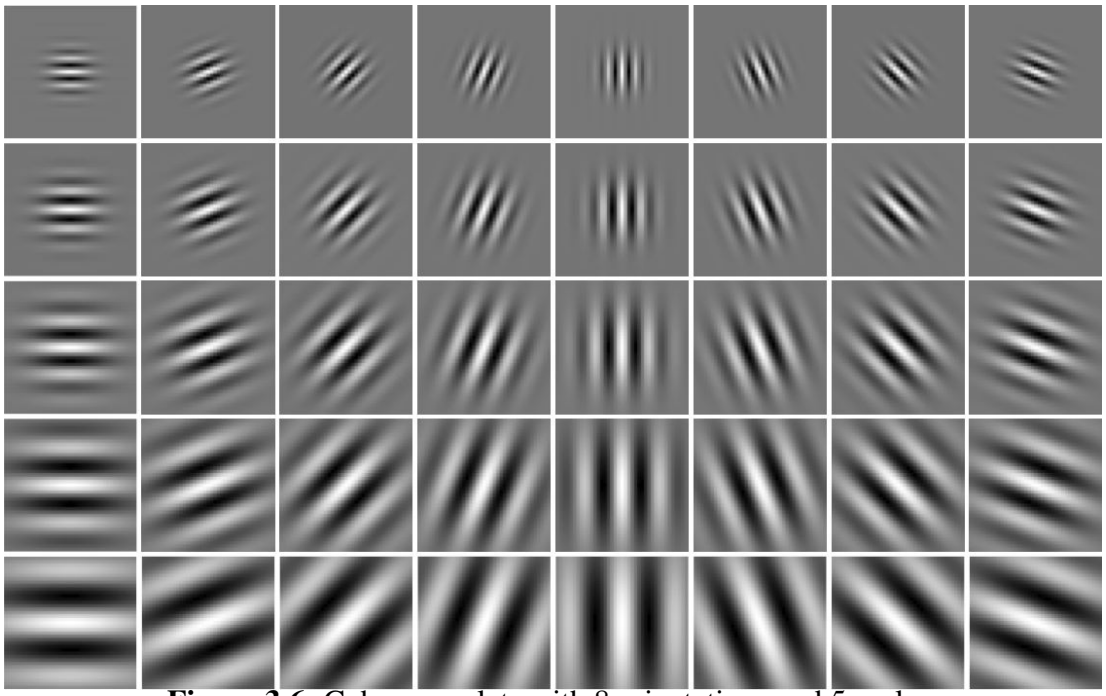


**Figure 3.6**: Gabor wavelets with 8 orientations and 5 scales.

**Spatial Augmentation:** Fisher vector encoding is efficient on feature space structure, however it does not consider spatial distribution of the features. Preceding to Fisher vector encoding, the spatial coordinates of the features are added to the local descriptors. The dense features can be represented as the following $[S_{xy}; \frac{x}{w} - \frac{1}{2}; \frac{y}{h} - \frac{1}{2}]$, where $S_{xy}$ are the PCA-SIFT features, $(x, y)$ is the center of patch and $(w, h)$ is the hight and the width of the patch.

In case when we use Gabor filters, spatial information and the corresponding scale and orientation are added to the local descriptors. The resulting local descriptor can be formulated as the following: $[G_{x,y}; x; y; S; O]$, where $G_{x,y}$ are the PCA-Gabor features of a block centered at $(x, y)$ which has $(S, O)$ scale and orientation.

**GMM Training and Fisher Vector Encoding:** The idea of Fisher vectors is to embed local descriptors by aggregating them into one single high dimensional vector which is more amenable to linear classification. Firstly, local features are used to drive a kernel from a generative model parameters. In this case, GMM is used by fitting its parameters into the the PCA-SIFT features. GMM can be referred as *probabilistic visual vocabulary*. The next step is encoding the gradient of the local descriptors' log-likelihood with respect to GMM parameters. GMM parameters are estimated on large set of local descriptors using the Expectation Maximization(EM) to optimize the log-likelihood. For Fisher vector computation the covariance of the GMM is assumed

to be diagonal and only the derivatives with respect to Gaussian mean and covariance are considered. This leads to a vectorial representation that captures the average first and the second order difference between dense features and each of GMM centers:

$$\Phi(k)^{(1)} = \frac{1}{N\sqrt{w_k}} \sum_{p=1}^{N} \alpha_p(k)(\frac{X_p - \mu_k}{\sigma_k}) \tag{3.5}$$

$$\Phi(k)^{(2)} = \frac{1}{N\sqrt{2w_k}} \sum_{p=1}^{N} \alpha_p(k)(\frac{(X_p - \mu_k)^2}{\sigma_k} - 1) \tag{3.6}$$

Where $w_k, \mu_k, \sigma_k$ are the GMMs' weights, means and diagonal covariance. $\alpha_p(k)$ is the soft assignment of p-th feature $x_p$ to the k-th Gaussian.

Fisher vectors dimensionality is $2Kd$ which depends on the number of the GMM $(K)$, and the dimensionality of SIFT features. The Fisher vector $\phi$ then is computed by stacking the differences ( the assignment of the local features to the first and second differences of GMM centers): $\phi = [\Phi(1)^{(1)}, \Phi(1)^{(2)}, ...., \Phi(K)^{(1)}, \Phi(K)^{(2)}]$. Fisher Vectors has many advantages: (i) it is generative to the data and discriminative, (ii) it can be computed using small number of parameters (GMM parameters), (iii) more importantly it has significant benefit that linear classifier can be learned efficiently using techniques such as Stochastic Gradient Descent.

### 3.4 Metric Learning

As a part of this thesis, high dimensional Fisher vectors are compressed into low dimensional and discriminative representation using metric learning. This compression is achieved by linear projection that serves two goals:

(i) Reducing the dimensionality of the Fisher vectors and making the method applicable on large scale datasets.

(ii) Improving the performance of recognition by making the projected subspace more discriminative with Euclidean distance such that positive and negative pairs are separated by a learned threshold within a margin in subspace.

#### 3.4.1 Large margin dimensionality reduction

The aim is to learn a Mahalanobis matrix $M = W^T W$ which is a convex formulation even in low-rank subspace. The learned matrix $(W \in R^{mxn}, m < n)$ projects high dimensional Fisher vector $\phi \in R^n$ to lower dimensional vector $W\phi \in R^m$ where $m$ is

the dimensionality of the subspace and $n$ is the original dimensions of Fisher Vectors. Such that the squared Euclidean distance: $d_w^2(\phi_i, \phi_j) = \|W\phi_i - W\phi_j\|_2^2$ of face images i and j is smaller than a learned threshold $b \in R$ if i and j belong to the same identity and larger if they belong to different people. As proposed in the paper [38] these conditions are imposed with a margin at least one as the following constraint:

$$y_{ij}(b - d_w^2(\phi_i, \phi_j)) > 1 \tag{3.7}$$

where $y_{ij}$ is 1 if the face images are for the same person and $-1$ otherwise. This idea can be shown in 3.7, we can see that the pairs are separated by margin with at least b learned threshold.

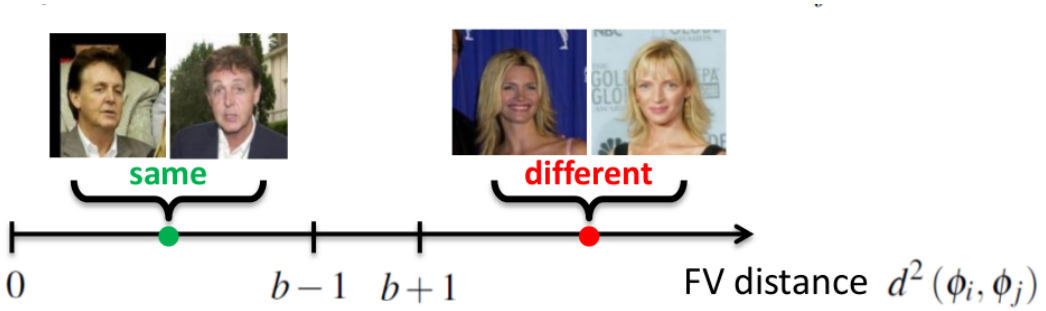



**Figure 3.7**: Positive and negative pairs are separated with a margin by metric learning [3].

The Euclidean distance in the lower subspace of m-dimensional can be seen as a low rank Mahalanobis distance in the original n-dimensional space.

$$d_w^2(\phi_i, \phi_j) = \|W\phi_i - W\phi_j\|_2^2 = (\phi_i - \phi_j)^T W^T W(\phi_i - \phi_j) \tag{3.8}$$

The direct optimization of $W$ is difficult due to the large number of $W$ parameters which can be over 2 billion for n = 67K dimensions of Fisher vectors. However, $W$ can can have $mn = 8.5M$ parameters if $m$ is set to 128 which is applicable when learning on large-scale datasets.

**PCA Whitening:** The initialization of $W$ is essential since objective function 3.9 is non-convex for $W$. In this work as suggested in the paper, $W$ is initialized with PCA-Whitening by choosing $m$ largest PCA dimensions. It differs from regular PCA by equalizing the magnitude of the dominant eigenvalues.

**Stochastic Sub-Gradient Descent (SGD):** Learning of $W$ is carried out by SGD, the following objective function is optimized during the learning process:

$$\underset{b,W}{\operatorname{argmin}} \sum_{i,j} max[1 - y_{ij}(b - (\phi_i - \phi_j)^T W^T W(\phi_i - \phi_j)), 0] \tag{3.9}$$

This problem is solved by using the stochastic sub-gradient method with number of iterations equal to $t = 1M$. In each iteration the algorithm take sample of pair face images $(i, j)$ such the the number of positive and negative pairs are balanced during the optimization process. The projection matrix is updated by

$$W_{t+1} = \begin{cases} W_t & \text{if } y_{ij}(b - d_w^2(\phi_i, \phi_j)) > 1 \\ W_t - \gamma y_{ij} W_t \psi & \text{otherwise} \end{cases} \tag{3.10}$$

where $\psi_{ij} = (\phi_i - \phi_j)(\phi_i - \phi_j)^T$ which is the outer product of the difference vectors. and $\gamma$ is the constant learning rate, which can be set or learned using validation set.

### 3.4.2 Low-rank joint metric learning

Joint Bayesian metric learning is a popular measurement of face similarity introduced by [15, 16] which is corresponding to the difference between low rank Mahalanobis distance $(\phi_i - \phi_j)^T W^T W(\phi_i - \phi_j)$ and a low-rank kernel(inner product) $\phi_i^T V^T V \phi_j$. This difference is used a score for face verification. The distance emerged from low-rank Mahalanobis metric measures the difference between pair images, while the score obtained from the inner product gives the similarity between pair images. For example, if the pair images belong to the same person, then we have small distance and large similarity score. The optimization for joint Bayesian is subjected to the objective function given by

$$\underset{W,V,b}{\operatorname{argmin}} \sum_{i,j} max[1 - y_{ij}(b - ((\phi_i - \phi_j)^T W^T W(\phi_i - \phi_j)) - \phi_i^T V^T V \phi_j), 0] \tag{3.11}$$

This differs from low rank Mahalanobis objective function 3.9 that it jointly add the inner product to the margin constraints since it is part of the score function as a similarity metric. Similarly to low rank Mahalanobis metric, the minimizer of 3.11 is found by stochastic sub-gradient descent method. The update of the $V$ and $W$ projection matrices is also achieved jointly in each iteration and computed as

$$W_{t+1} = \begin{cases} W_t & \text{if } y_{ij}(b - (d_w^2(\phi_i, \phi_j) - (V\phi_i)^T (V\phi_j)) > 1 \\ W_t - y_i j(\gamma d_w^2(\phi_i, \phi_j)(\phi_i - \phi_j)) & \text{otherwise} \end{cases} \tag{3.12}$$

$$V_{t+1} = \begin{cases} V_t & \text{if } y_{ij}(b - (d_w^2(\phi_i, \phi_j) - (V\phi_i)^T(V\phi_j)) > 1 \\ V_t + y_i j((\gamma V \phi_i)\phi_j^T + (\gamma V \phi_j)\phi_i^T) & \text{otherwise} \end{cases} \tag{3.13}$$

Where in each iteration the pairs $(\phi_i.\phi_j)$ can be positive or negative, and according to this $y_{ij}$ value is 1 if the pairs belong to the same person and $-1$ otherwise. $\gamma$ is a learning rate that can be set manually or trained by validation data.

The threshold $b$ we learn in metric learning that separate positive and negative pair images is updated in both low-rank Mahalanobis and joint metric learning as following:

$$b = b + y_{ij} * \text{ gamma rate} \tag{3.14}$$

Where $y_{ij}$ is 1 if the training pair images belong to the same person, and $-1$ otherwise. Gamma rate is the learning rate of $\gamma$ that can be set previously.

### 3.4.3 Diagonal metric learning

In diagonal metric learning the learned $W$ projection matrix is assumed to be diagonal. Learning a diagonal Mahalanobis matrix is equivalent to learning weights to the features and evaluating the Euclidean distance with the re-weighted dimensions. the learning of diagonal matrix is carried by using the conventional linear SVM formulation, where the features are the vectors of squared difference between the compared Fisher vectors. The learning optimization is subjected to the objective function given by

$$\underset{u_k \geqslant 0}{\operatorname{argmin}} \sum_{i,j} max[1 - y_{ij}(b - d_u^2(\phi_i - \phi_j)), 0] \tag{3.15}$$

Where $d_u^2(\phi_i - \phi_j)) = \sum_k u_k(\phi_i - \phi_j)^2$ and the projection matrix is updated as shown in equation 3.16

$$W_{t+1} = \begin{cases} W_t(1 - \gamma\lambda) - (\gamma X_t) & \text{if } W^T X > -1 \\ W_t(1 - \gamma\lambda) & \text{otherwise} \end{cases} \tag{3.16}$$

In 3.16, $W$ is the diagonal projection matrix, $t$ is the iteration of sub-gradient stochastic descent, gamma and lambda are the learning rates. $X_t$ is the difference between squared distance of positive and negative pairs taken in each iteration. The difference is given

in equation 3.18 and $\gamma$ is updated in each iteration by

$$\gamma_{t+1} = 1(t\lambda_t) \tag{3.17}$$

$$X = (\phi_1^i - \phi_2^i)^2 - (\phi^x - \phi^y)^2 \tag{3.18}$$

In 3.18 $(\phi_1^i, \phi_2^i)$ are the Fisher vectors of positive pairs, and $(\phi^x, \phi^y)$ are the Fisher vectors of pairs belong to different identities.

### 3.5 Linear Support Vector Machines (SVMs)

In this work we use linear SVM for face track recognition in Accio dataset. Support Vector Machines are one of the most prominent machine learning algorithms. It is has been widely and successfully used in applications such as face recognition, text classification ... ets. In SVMs, typically we have the labeled training data $\{X_i, y_i\}$, where $i = 1, ...l$ and $y_i \in \{-1, 1\}$. $x_i$ has a dimension equal to $d$, $x_i \in R^d$. The aim is to find a hyperplane that separate the positive and the negative examples such that the margin between them is maximized. In linear SVMs which is the basic one, it is assumed that the data can be separated perfectly. Training time of SVM depends on the number of example not the dimension of the features which is an important advantage. SVMs optimize the projection matrix $W$ by formulating the problem as constrained optimization problem such that: we minimize $\{\|W\|\}^2$ subjected to $\{y_i(W^T X_i + b) >= 1\}$ where the samples on the right of the hyperplane has value > 0 or 1 otherwise. This problem is known as quadratic optimization problem where it has only one global minimum and can be efficiently solved by optimizer such as Lagrange multipliers.

During the evaluation within and cross movies face track recognition, we use binary linear SVMS (one vs. all) technique. For example in Accio dataset, we obtain Linear SVM model for each character separately by training each character model with its face tracks as positive samples and the rest of characters face tracks as negative samples. Then we use this generated linear SVM model for testing in the same behavior.

### 3.6 Performance Evaluation Methods

In this section we explain the performance measurement techniques we use for face track retrieval, recognition and face verification.

**Mean Average Precision:** For face track retrieval, mean average precision is the main method to assess the performance of the within and across movies retrieval tasks in Accio dataset.

**Precision:** precision is the percentage of the retrieved face tracks that are relevant.

$$P = \frac{\text{\#of relevant face tracks}}{\text{\#of face tracks retrieved}} \tag{3.19}$$

During the evaluation on Accio dataset, for a face track query the system returns a ranked sequence of face tracks based on the distance between the face track query and the database, Average Precision calculates the value of $p(q)$ in 3.21 by summing over each position in the returned ranked sequence of face tracks and dividing by the number of retrieved elements (face track database size).

$$AP = \frac{\sum_{k=1}^{n}(P(k)\times rel(k))}{\text{number of relevant tracks}} \tag{3.20}$$

Mean Average Precision (mAP)score for an evaluation with a set of face tracks queries and database is the mean of the average precision scores for each face track query.

$$mAP = \frac{\sum_{q=1}^{N} AP(q)}{N} \tag{3.21}$$

In addition to mAP, we report the results in term of precision @k. Since face track queries may have hundreds of relevant tracks in database, we decide that it is meaningful to check the precision at certain k values correspond to the number of first k relevant results in the ranked retrieved face tracks. For example precision @10 means that we look to the precision among the 10 retrieved results. We check precision @k for the following k values: $[1,5,10,20,50,100]$. It is important to mention that, if the query does not have k relevant face tracks in the database, its result is ignored and not included in the performance. Sections 4.2.2 and 4.3 describe the results of face track retrieval in term of mAP and precision @k and compare them using different metric learning with variant scenarios.

**Accuracy:** In face track recognition, accuracy corresponds to the number of correctly classified face tracks according to the model. We have the ground truth identities for most of face tracks. Accuracy is the rate of true positive (tp) plus true negative(tn) face tracks to the total number of face tracks. Section 4.4 discuss the results of face track recognition.

$$Accuracy = \frac{\text{tp} + \text{tn}}{\text{total face tracks}} \tag{3.22}$$

### 3.6.1 Receiving operator characteristic (ROC) curve:

ROC analysis of a performance is one of the most popular method used in machine learning and computer vision. ROC is a graphical plot that shows the performance of a binary classifier with variant threshold. The plot created by plotting the true positive rate (TPR) vs. false positive rate (FPR).

**Area Under The Curve (AUC):** AUC is a popular metric for classification. It consider various thresholds values which result in different TPR and FPR. When the threshold is increased, we get higher value of TPR, and higher value of FPR otherwise.

**Equal Error Rate (EER):** ROC-EER is the accuracy at ROC operating point where TPR and FPR are equal. Results of Fisher vector analysis on face verification task are refereed in term of ROC-AUC and ROC-EER. All the results are given in section 5.2.

# 4. FACE TRACK RETRIEVAL AND RECOGNITION ACROSS AGE

In this chapter we discuss the system of face track retrieval and recognition across age. Firstly we introduce the novel dataset we use during our study: Accio. We provide important statistics and details about the dataset. A part from the dataset, we explain scenarios for face track retrieval and recognition. Then finally, experiment's sections we provide in detail the baseline and improved results by using the reduced dimension of face tracks' Fisher vectors and multiple metric learnings.

## 4.1 Accio: Harry Potter Movies Aging Dataset

Available datasets for age progression are constrained to still images based datasets. As a results there is not a study which considers age progression within the scope of videos. Due to the lack of public video based dataset that address aging factor in face recognition and regarding the fact that data research and development require data we introduce Accio dataset as a novel dataset to study and develop age invariant face recognition and verification methods. Face recognition on video is always harder than still images based face recognition due to various challenges such as variant illumination, different face pose and expression within a video shot.

### 4.1.1 Data source

This dataset is collected and organized using the eight Harry Potter movies that were released in a period of ten years – 2001 to 2011. Each movie in itself contains multiple interesting challenges (i) *illumination* - large number of bright and dark scenes; (ii) *pose* - many non-frontal tracks, actors typically do not look at cameras; (iii) *resolution* - tracks from 25 to 500 pixels; and (iv) *distractors* - roughly 40% of all face tracks are *unknown* characters which appear in the background making the problem an open set task. Finally, the use of all eight movies acts as a good source of variation in the facial appearance due to *aging* and provide a good study case to measure the performance of face recognition system against aging progression.

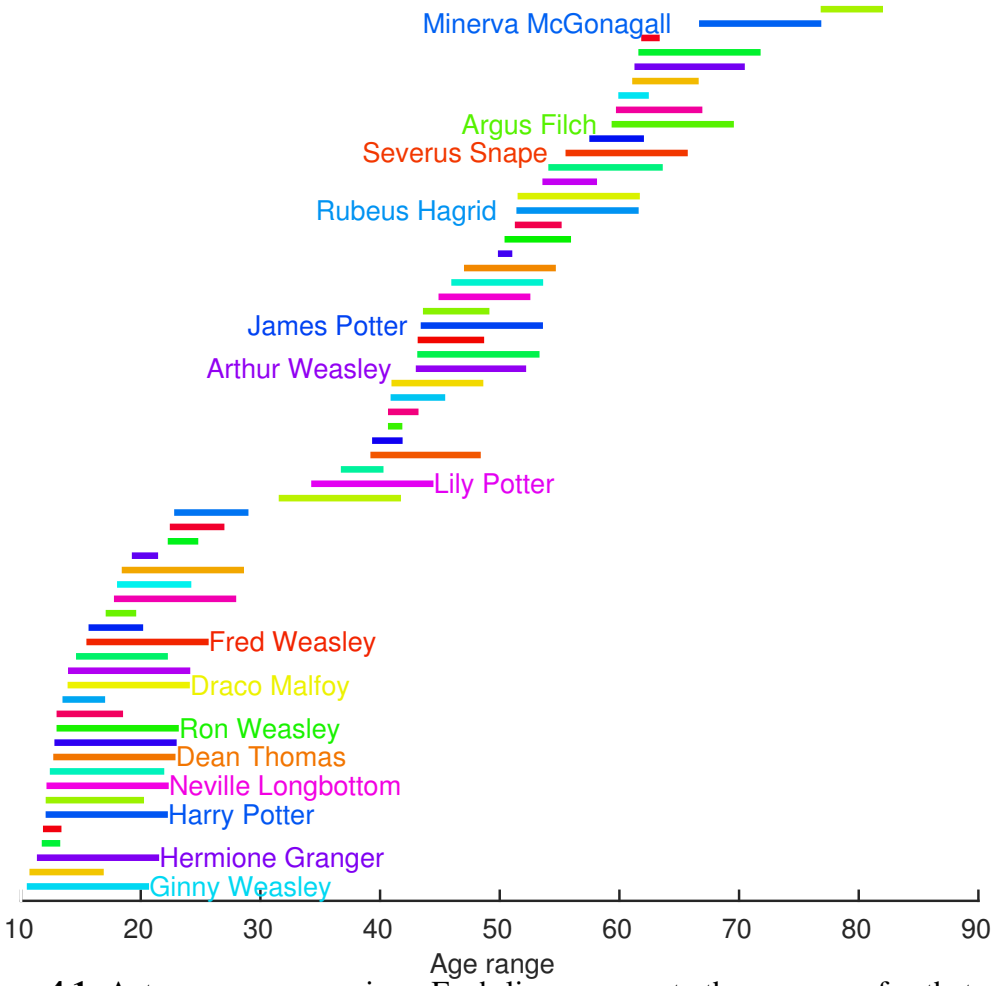



**Figure 4.1**: Actors age progression. Each line represents the age span for that actor appearing in the movies.

To obtain the face tracks, we first detect all Harry Potter videos shots by Displaced Frame Difference (DFD) we explain in section 3.1. Within each shot, following [6] we perform tracking-by-detection using a particle filter. This system is described in section 3.2. We use multi-pose detectors which cover the entire range of *pan* and *roll* face angles. Face tracks are labeled manually and named as the characters appear in Harry Potter movies. In addition, actor name is included in annotation since some characters are changed through eight movies.

In Harry Potter movies we can see how the facial appearance of characters changes through movies. This an important property for studying age invariant face track retrieval and recognition. Figures in 4.2 provide faces of some young characters in the first, fourth and last movie which apparently prove the facial changes through years. The dataset contains a large number of tracks of young actors (age < 20). We believe that this is an important aspect since most of the changes in facial appearance occur

early in life. Fig. 4.1 presents the age variation for multiple actors. Please note that the figure plots only those actors which appear in at least 2 movies.

(a) Harry (age 12, 16, 22) (b) Ron (age 13, 17, 23)

(c) Hermione (age 11, 15, 21)

**Figure 4.2**: Change in character appearance across the movies.

### 4.1.2 Face track descriptors

For face track representation we adopt the usage of Fisher vector pipeline explained in section 3.3.1. Face tracks of Accio dataset are encoded using this state-of-the-art Video Fisher Vector Faces ($VF^2$) descriptor [8]. $VF^2$ is a face track representation that aggregates a large set of dense SIFT features. SIFT features are pooled from the entire face track and augmented with their normalized $\{x,y\}$locations. We follow the steps and parameters from [3, 8] and use Root-SIFT with a step of 1 pixel and 5 scales with a factor set to $\sqrt{2}$. The resulting 128 dimensional feature is down-projected via PCA to 64 dimensions and augmented with the spatial location $(x,y)$. A Gaussian Mixture Model (GMM) of 128 Gaussians is trained on all the face tracks and Fisher encoding is performed. This yields a $2 \times 128 \times 66 = 16896$ dimensional descriptor for each track. To compare two tracks we use the standard Euclidean distance between two Fisher vector descriptors for our baseline experiments.

### 4.1.3 Statistics

Table 4.1 presents the number of characters, the number of tracks in each movie and a split between *named* and *unknown* (distractor) tracks.The dataset contains in total of 38,464 face tracks. 22,830 (59.4%) of these face tracks are labeled with one of 121 character identities as they appear in the film series. The rest of the face tracks (40.6%) belong to un-named cast and are treated as distractors for the retrieval evaluation.

Each movie in the Harry Potter film series has varying number of characters and duration. In consequence, each movie has different number of face tracks and ground truth identities. The three main characters (*Harry Potter*, *Hermione Granger*, and *Ron Weasley*) encompass roughly 40-60% of all face tracks in each movie.

**Table 4.1**: Number of tracks and characters across the movies.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 | Total |
|---|---|---|---|---|---|---|---|---|---|
| # characters | 36 | 42 | 34 | 44 | 47 | 41 | 56 | 54 | 121 |
| # face tracks | 5,249 | 5,335 | 3,919 | 7,616 | 5,850 | 3,354 | 2,910 | 4,231 | 38,464 |
| # unknown tracks | 2,006 | 1,874 | 1,437 | 4,237 | 2,316 | 1,116 | 623 | 2,025 | 15,634 |
| # named tracks | 3,243 | 3,461 | 2,482 | 3,379 | 3,534 | 2,238 | 2287 | 2,206 | 22,830 |

labeled face tracks span a range of age between 10 to 88 years. Note that we compute the age based on the movie release date and the birth year, which can add a slight discrepancy up to a couple years. The first movie was released in Nov. 2001, and the last in Jul. 2011. The largest age difference between face tracks for the same actor is 10 years for actors which appear in both the first and last movie.

Table 4.2 shows the distribution of tracks and number of characters across various age spans. We see an equitable distribution of characters among the various age groups. However, most of the tracks belong to the first 10-19 (*Young*) age range. We notice how young face tracks are more dominant in Accio dataset. Young and kids individuals are more difficult to recognize than older ones, since facial appearance changes in early stage of age progression are more than older stages. This shows how harry potter movies are good case to study the impact of age progression.

**Table 4.2**: Actors age distribution.

| | 10-19 | 20-39 | 40-49 | 50+ | Young | Adults |
|---|---|---|---|---|---|---|
| # characters | 33 | 34 | 26 | 44 | 33 | 92 |
| # labeled tracks | 12,598 | 4,303 | 1,728 | 3,636 | 12,598 | 9,667 |

Finally, we present in Table 4.3 a comparison between existing datasets. Note that unlike other datasets, ours works with face tracks, which on average are 50 frames (2 seconds) long thus yielding over 1.9 million face images.

**Table 4.3**: Comparison of age-invariant face recognition datasets.

| | video? | #images | #people | Age span |
|---|---|---|---|---|
| FGNET [10] | No | 1,002 | 82 | 0-45 |
| MORPH [11] | No | 55,134 | 13,618 | 0-5 |
| CACD [12] | No | 163,446 | 2,000 | 0-10 |
| Accio (Ours) | Yes | 38,464 tracks | 121 | 0-10 |

## 4.2 Experiments

We first define the evaluation protocols, followed by baseline experiments for face track retrieval. Then we introduce the baseline and improved results using multiple metric learning algorithms with different scenarios. Some experiments are carried out by using the full dimension of Fisher vectors, while the rest are achieved after applying PCA dimensionality reduction to face tracks' local descriptors. We compare the improved results to the baseline in each scenario to check the impact of metric learning and the difference between the baseline and improved results. The retrieval experiments are evaluated using two popular measures: (i) mean average precision (MAP), and (ii) precision @ $k$ (number of correct results in top $k$) as we explained in 3.6.

### 4.2.1 Evaluation protocol

**Tasks:** The face track retrieval is performed as two tasks:

(i) *Within* movies: In this setting, we look at individual movies one at a time. Each named track in the movie is used as a query, while the remainder (other named tracks + unknown tracks) from the same movie form the database. This setting evaluates the retrieval performance subject to typical face recognition challenges such as variant pose and illumination.

(ii) *across* movies: In this setting we analyze the effect of age variation among the actors by evaluating face track retrieval across movies. We consider each named track within one movie (for example HP-1 ) as a query, while all tracks of a different movie (for example HP-2 ) form the database.

**Benchmarks** We suggest two protocols which involve both the above tasks, however vary in the extent to which data can be used to learn supervised or unsupervised models.

(i) *Restricted*: In this protocol the experiments should not use any external data. That is, only the query track is available as a positive training sample and retrieval models should not be trained using other *Accio* face tracks, or any external face images/tracks. This can be used for example to compare different face track descriptors, or perform automatic and unsupervised methods such as query expansion and domain adaptation.

(ii) *Unrestricted*: In this protocol the researchers are allowed to use *external* data for training models, for example to model the changes in facial appearance through aging.However, no other tracks from the *Accio* dataset (except the query track) are to be used for training actor models.

Finally, if the data is used in a different manner that does not comply with either of the two previous protocols, for example the face tracks can be used for a classification task [7], the researchers can refer to this as a *free-for-all* scheme.

In this work, we use the last protocol *free-for-all* scheme to carry out the experiments of metric learning on Accio dataset. We apply two type of metric learning according to the usage of face tracks. First type uses a percent of face tracks from all movies for training a metric model and we refer to it as *metric learning on all movies data*. The second type uses percent of face tracks from pair movies such as first and second movies, to obtain metric learning and then we test on the rest of face tracks of the two movies using the generated metric learning model. We refer to the second type of metric learning as *metric learning on pairwise movies*.

### 4.2.2 Baseline evaluation on all face tracks

This part of experiments provide the experimental results of the face track retrieval tasks. we use the full dimensions of face track descriptors. In addition we apply the evaluation on the whole face tracks. The aim here is to show the impact of age progression across movies.

**Within movies face track retrieval**: We now present the results of within movie face track retrieval (Task 1) under the restricted protocol. In each movie, the labeled tracks

are used as a query, while all tracks from the same movie form the database. For example, in HP-1, we have 3243 tracks (taken one at a time) as query and 5248 tracks (the number of tracks minus the query) as part of the database.

Table 4.4 presents the performance of retrieval on the eight movies. Precision @ $k$ (p @ $k$) values are presented in multiple steps [1, 5, 10, 20, 50, 100], while the MAP is used as the overall performance indicator. As expected the performance for small values of $k$ is much higher and this can be used to perform query expansion.

**Table 4.4**: Within movies evaluation.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| p @ 1 | 90.8 | 89.3 | 90.9 | 85.1 | 89.4 | 91.1 | 91.3 | 89.4 |
| p @ 5 | 81.5 | 79.0 | 79.1 | 72.3 | 79.3 | 80.4 | 82.1 | 78.6 |
| p @ 10 | 75.9 | 71.6 | 70.2 | 64.5 | 72.0 | 72.2 | 75.3 | 69.7 |
| p @ 20 | 71.9 | 64.0 | 61.5 | 57.3 | 63.7 | 64.3 | 70.7 | 63.0 |
| p @ 50 | 67.9 | 57.6 | 53.0 | 54.7 | 59.4 | 55.8 | 68.0 | 57.2 |
| p @ 100 | 70.9 | 54.8 | 53.1 | 58.5 | 55.5 | 51.1 | 62.8 | 53.0 |
| MAP | 42.40 | 31.73 | 31.19 | 28.36 | 32.09 | 30.38 | 38.55 | 33.21 |

**Table 4.5**: across movies face track retrieval performance (MAP).

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 42.0 | 33.7 | 28.4 | 26.9 | 23.5 | 24.6 | 28.7 | 21.9 |
| HP-2 | 32.4 | 31.7 | 24.2 | 22.1 | 20.2 | 21.5 | 24.4 | 19.6 |
| HP-3 | 24.1 | 21.7 | 30.8 | 20.8 | 19.1 | 21.7 | 24.1 | 17.1 |
| HP-4 | 28.9 | 24.9 | 28.0 | 28.9 | 22.4 | 24.3 | 25.4 | 21.8 |
| HP-5 | 21.5 | 19.7 | 22.8 | 18.6 | 32.1 | 23.1 | 26.4 | 21.9 |
| HP-6 | 23.1 | 19.8 | 25.3 | 21.3 | 24.9 | 30.4 | 25.2 | 19.0 |
| HP-7 | 25.7 | 22.3 | 25.7 | 21.0 | 27.7 | 25.8 | 38.5 | 26.5 |
| HP-8 | 20.8 | 19.9 | 21.0 | 19.1 | 22.7 | 21.6 | 28.7 | 33.4 |

**Across movies face track retrieval**: While the previous section evaluated performance through challenges of pose and illumination, we now look primarily at the problem of aging (Task 2). We expect that the retrieval performance is good to fair when the age difference is small (within same movie, or $\pm$ 1 movie), while it deteriorates as we compare across a large time gap (for example HP-1 vs. HP-8).

Table 4.5 presents overall retrieval performance in MAP comparing all eight movies. Rows represent the movie from which query tracks are obtained, while the tracks of the movie in columns serves as the database (for example query from HP-1 and database

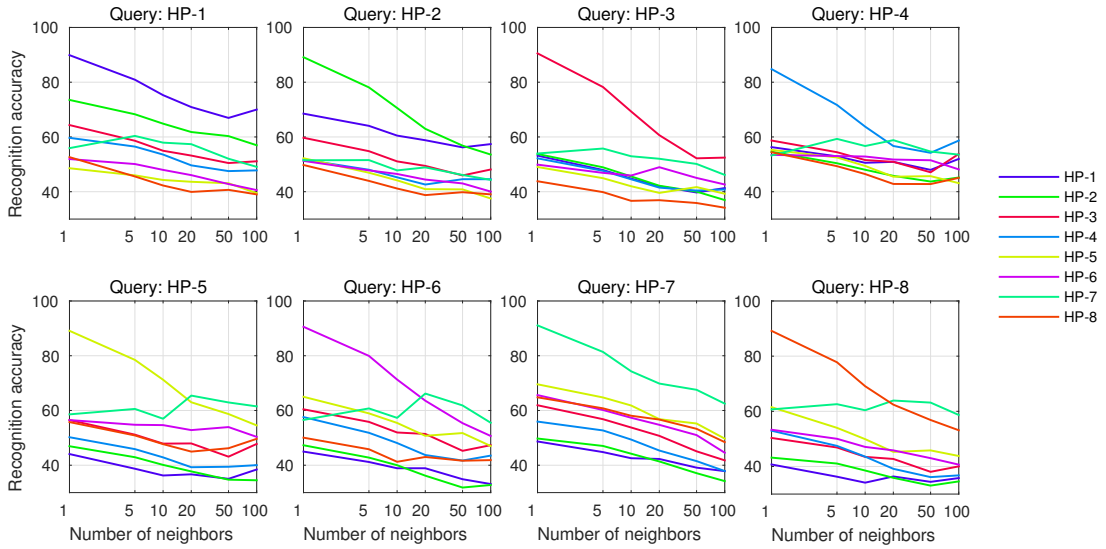



**Figure 4.3**: Precision @ $k$ results for across movies retrieval. This figure is best viewed in color.

from HP-7 shows a MAP of 27.4). As can be seen in the table, the larger the age gap between the query and database the MAP performance reduces. HP-7 appears to be an exception and this could be attributed to the fact that this movie has the least number of tracks, many ($\sim$60%) of which belong to the three primary characters. The diagonal represents within movie performance and is the same as the last row of Table 4.4.

In Table 4.6, we consider all queries from the first movie HP-1 and split them into two based on age. Tracks of young actors (age < 20) when used as queries tend to perform worse in the retrieval across movies as compared to tracks of adults. Empirically we see that the face appearance changes more during the youth years and thus has a stronger influence (more errors) on the retrieval performance.

Finally, we present the precision @ $k$ scores for each movie in Fig. 4.3. Each subplot represents queries from a single movie (mentioned in the title) and the databases are all movies (plotted as separate lines). For any subplot, note how the precision is higher for movies around the movie from which the query tracks are selected. For example, choosing queries from HP-1, the best performance is seen when comparing against tracks from HP-2, HP-3, HP-4. Similarly, for queries from HP-2, the performance against tracks from HP-1 and HP-3 is better than other movies. There are some exceptions, specially at smaller values of $k$, however the general trend is fairly clear. Together with Table 4.5, the analysis is a clear indication that age variation is a strongly felt effect in the dataset.

**Table 4.6**: Comparison of MAP scores for adult vs. young actors. The queries are taken here from HP-1 movie.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| Adult | 47.5 | 40.9 | 33.6 | 32.5 | 30.5 | 29.8 | 23.0 | 21.8 |
| Young | 40.9 | 31.2 | 25.2 | 22.9 | 19.3 | 20.8 | 28.4 | 19.6 |

## 4.3 Improving Baseline Face Track Retrieval Results

Following the previous retrieval two tasks' experiments, we conduct multiple experiments to improve the baseline results of retrieval using metric learning. The goal is to reduce the gap between the retrieval results of two different movies in across movies evaluation. We carry out experiments with metric learning methods we explain in section 3.4. We apply two different scenarios: metric learning on pairwise movies and a metric learning on data pooled from all movies. In the following section we explain the two experiments settings and provide the retrieval results. We apply PCA dimensionality reduction on Fisher vectors to decrease the dimensions of Fisher vectors from 16,896 to 1000 since PCA helps to highlight the similarity of features and the importance of reduced dimensions. In addition, it helps to accelerate metric learning since we have less number of dimensions.

### 4.3.1 Metric learning on pairwise movies

In this experiment, for a pair of different movies we use 10% of movies face tracks to train a metric model and use the projection matrix obtained from the metric model for the retrieval on the rest of face tracks. The evaluation is achieved by the Euclidean distance between the two movies' face tracks. The learning is applied by combining the training face tracks' data from both movies and generate positive and negative pairs from the face tracks ground truth identities to use for for training stage. Table 4.7 illustrates the number of positive and negative pairs generated when using the combined face tracks from both first and second movies. The table shows different percentage values for training data from both movies which generate variant number of training pairs.

**Table 4.7**: Comparison of number of pairs when using different percent of face tracks from pair movies.

| training percent | 1% | 2% | 5% | 10% | 20% |
|---|---|---|---|---|---|
| #positive pairs | 302 | 1173 | 7263 | 28621 | 113709 |
| #negative pairs | 6139 | 14227 | 62862 | 222365 | 829542 |

Preceding to the evaluation of pairwise metric learning across movies, we obtain across and within movies retrieval results using the Euclidean distance without any learning and using the reduced dimensions of Fisher vectors. This experiment gives the results similar to the baseline results in section 4.2.2. However, the retrieval is made only on the test face tracks using the Euclidean distance.

**Table 4.8**: Pairwise baseline's cross movies FT retrieval: mAP.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 39.4 | 31.0 | 23.7 | 22.5 | 18.9 | 20.1 | 27.8 | 16.9 |
| HP-2 | 29.9 | 29.3 | 20.2 | 18.3 | 16.3 | 18.4 | 23.4 | 15.9 |
| HP-3 | 20.7 | 19.0 | 27.3 | 17.3 | 15.5 | 18.9 | 22.8 | 13.5 |
| HP-4 | 25.7 | 22.0 | 24.2 | 26.4 | 18.0 | 21.4 | 24.9 | 17.4 |
| HP-5 | 19.0 | 17.3 | 19.5 | 15.8 | 28.6 | 20.9 | 26.3 | 18.2 |
| HP-6 | 19.3 | 16.9 | 21.2 | 17.2 | 19.9 | 27.2 | 24.5 | 15.3 |
| HP-7 | 22.8 | 19.6 | 21.1 | 17.1 | 23.2 | 22.1 | 35.7 | 22.8 |
| HP-8 | 18.5 | 17.7 | 17.2 | 16.0 | 18.8 | 19.3 | 27.5 | 29.8 |

In addition, in *task 1* when the retrieval is done within the same movie, we use 20% of the face tracks for training the metric learning. Since we think that it would be fair comparison to improvement of across movies retrieval. Furthermore, there are two reasons for this selection:

(i) The face tracks' number is closer to the one we have for pairwise metric learning across movies.

(ii) The number of generated positive and negative pairs for metric learning training is also comparable to the ones in across movies pairwise metric learning.

Table 4.8 shows the mAP results of face track retrieval using the reduced dimensions of Fisher vectors and Euclidean distances. Diagonal elements of the table shows the mAP of within movies retrieval *(task 1)* where evaluation is subjected to the standard challenges such as pose and illumination, and the rest of elements show the mAP results of across movies retrieval *(task 2)* where time gap and facial changes exist

due to age progression. We can see that the diagonal elements have higher values, nonetheless the performance degrades when we go farther from diagonal elements. This show the impact of aging factor in addition to secondary technical challenges such as video quality and illumination changes across movies for *task 2*

Figure 4.4 presents the precision @$k$ scores for across movies face classification. Each sub-figure represents different movie which used for query mentioned in its title. Lines in figures represent the movies. Note that when query movies is near to the database movie, the precision is better which indicates the effect of aging due to the years gap between movies release dates. In addition there is a decrease in precision when $k$ values are higher. For example, when query is from HP-1 the best results are obtained from comparing to HP-1, HP-2 and HP-4 databases. Queries of other movies have similar trend in their results. This case clarify the impact of years' gap between movies on the performance.

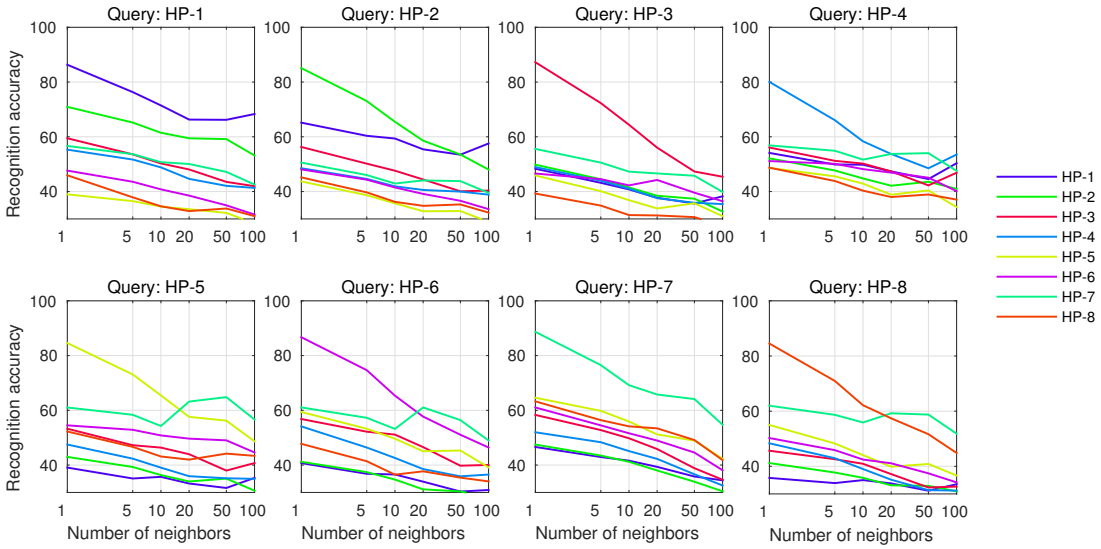



**Figure 4.4**: Baseline pairwise precision @ $k$ results for across movies retrieval. This figure is best viewed in color.

**Joint Metric Learning:** in this experiment we apply Joint metric learning as explained in 3.4.2 by training the model on 10% of the face tracks from two movies. The projection matrices $(W,V)$ with a dimension equal to 128 are trained using the negative and positive pairs of the training data. Table 4.9 shows the mAP results of withing and across movies retrieval. Compared to the baseline results of using the reduced dimensional Fisher vectors and without learning, joint metric learning results in huge improvement in term of mAP for both tasks: within and across movies. This metric

learning leads to a big improvement in mAP such that the average improvement in reference to baseline mAP of across and within movies retrieval is 40.1%. Most of the progress in the performance occur in *task 2* where the mAP results of baseline is usually low compared to *task 1* results. The improvement indicates the impact of using this metric learning and its ability to capture the similarity between movies face tracks. An example of improvement in across movies' mAP, the mAP improvement between HP-1 and HP-2 retrieval reaches 47.3%. This improvement is due to the supervised pairwise learning between movies. The metric learning can learn the facial changes due to the years' gap between face tracks and try to impose the similarity and the difference according to the given positive and negative face tracks.

**Table 4.9**: Pairwise joint metric's cross movies FT retrieval: mAP.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 69.5 | 66.9 | 64.4 | 64.1 | 64.0 | 67.1 | 73.7 | 64.2 |
| HP-2 | 67.1 | 65.2 | 60.6 | 57.9 | 59.7 | 62.3 | 66.6 | 57.3 |
| HP-3 | 63.5 | 61.1 | 61.2 | 54.0 | 53.8 | 56.1 | 62.9 | 52.1 |
| HP-4 | 63.2 | 60.9 | 59.6 | 54.2 | 56.5 | 58.1 | 65.5 | 58.8 |
| HP-5 | 60.4 | 58.1 | 58.0 | 54.4 | 62.7 | 56.9 | 63.4 | 54.4 |
| HP-6 | 66.7 | 63.3 | 59.8 | 58.7 | 63.0 | 64.9 | 63.4 | 53.0 |
| HP-7 | 73.2 | 68.6 | 68.1 | 65.5 | 68.6 | 68.1 | 67.1 | 67.2 |
| HP-8 | 58.9 | 59.8 | 56.5 | 55.4 | 58.5 | 55.9 | 64.8 | 60.2 |

In addition, figure 4.5 give the precision @$k$ results after pairwise joint metric learning. The figure points out how the gap between lines which represent the movies of queries and database is reduced. In subgraphs the title gives the movie which is used for face tracks queries. In addition there is an improvement in term of precision too for all k values as we can observe in table 4.10. The table presents the average difference of precision @k between the baseline and the joint metric learning results.

**Table 4.10**: The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values.

| k values | @1 | @2 | @10 | @20 | @50 | @100 |
|---|---|---|---|---|---|---|
| Improvement Percentage | 21.9 | 26.3 | 29.3 | 33.5 | 41.4 | 47.6 |

**Joint Metric Learning on Full dimensional Fisher Vectors:** Previous experiment presents the results we obtain after applying dimensionality reduction by PCA. In this

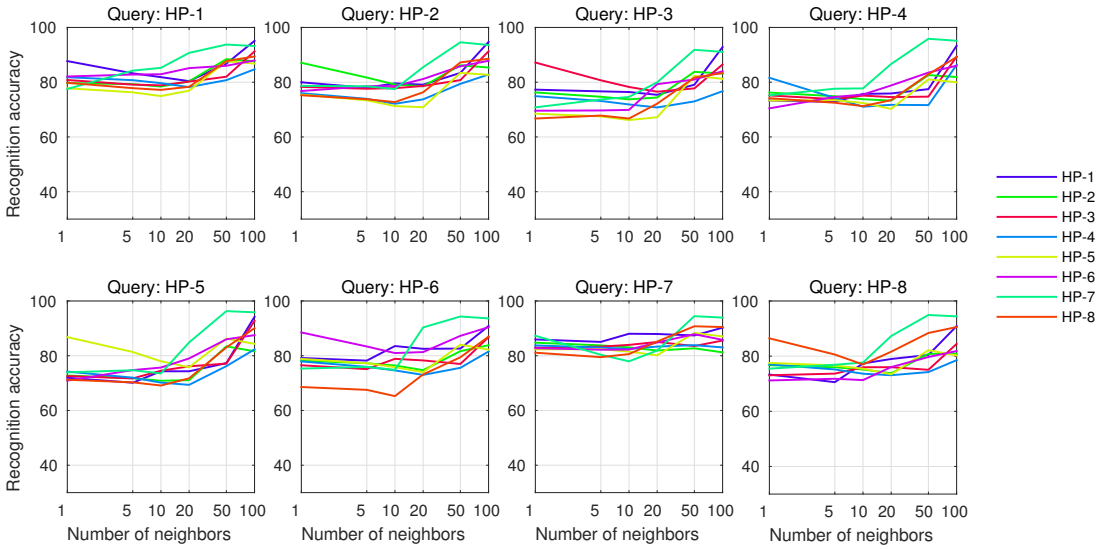



**Figure 4.5**: Pairwise joint metric learning's precision @ $k$ results for across movies retrieval. This figure is best viewed in color.

part we introduce the results of face track retrieval using the original dimensions of face track descriptors. Table 4.11 gives the mAP results of both retrieval tasks (1 and 2). Compared to mAP results in table 4.9, we can see that the results of joint metric learning after appyling PCA is better. For example the average improvement of mAP after using joint metric learning on full dimensional fisher vectors is 33.9%. However, the gain in mAP results is higher when we apply joint metric learning on reduced face track descriptors such that the improvement is 40.1%. Figure 4.6 also gives the results of face classification using precision @k. We can see also that the results in figure 4.5 are more consistent and stable such that the gap between lines is reduced more when we apply joint metric learning on reduced face track's descriptors. In addition, table 4.12 gives the average differences of precision @k between baseline and joint metric learning results. We can notice that the improvement is less than the progress we obtain in table 4.10, which emerges from applying joint metric learning or reduced Fisher vectors.

This experiment show the benefit of reducing the dimensionality of face track descriptors by PCA such that it helps to highlight the similarity of features and the importance of reduced dimensions to further improve the results of retrieval tasks.

**Low-rank Mahalanobis metric learning:** The second metric learning method we adopt for this experiment is low-rank Mahalanobis metric learning as explained in 3.4.1 section. The dimensions of the fisher vectors were reduced from 1000 to 128

**Table 4.11**: Cross movies FT retrieval: mAP.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 62.7 | 64.6 | 61.6 | 60.6 | 59.8 | 63.9 | 72.0 | 60.2 |
| HP-2 | 64.4 | 53.7 | 58.7 | 55.1 | 57.9 | 58.7 | 63.3 | 55.1 |
| HP-3 | 61.1 | 58.6 | 48.2 | 52.5 | 52.5 | 52.7 | 63.0 | 50.6 |
| HP-4 | 60.6 | 58.1 | 57.1 | 45.3 | 54.0 | 56.5 | 64.1 | 56.9 |
| HP-5 | 58.0 | 55.7 | 55.8 | 51.2 | 52.3 | 55.6 | 59.9 | 52.2 |
| HP-6 | 63.3 | 60.3 | 57.9 | 56.6 | 61.3 | 60.8 | 60.7 | 49.2 |
| HP-7 | 71.5 | 65.6 | 68.1 | 62.7 | 64.8 | 66.8 | 64.5 | 62.5 |
| HP-8 | 55.6 | 56.1 | 54.6 | 54.1 | 56.1 | 52.2 | 61.6 | 57.2 |

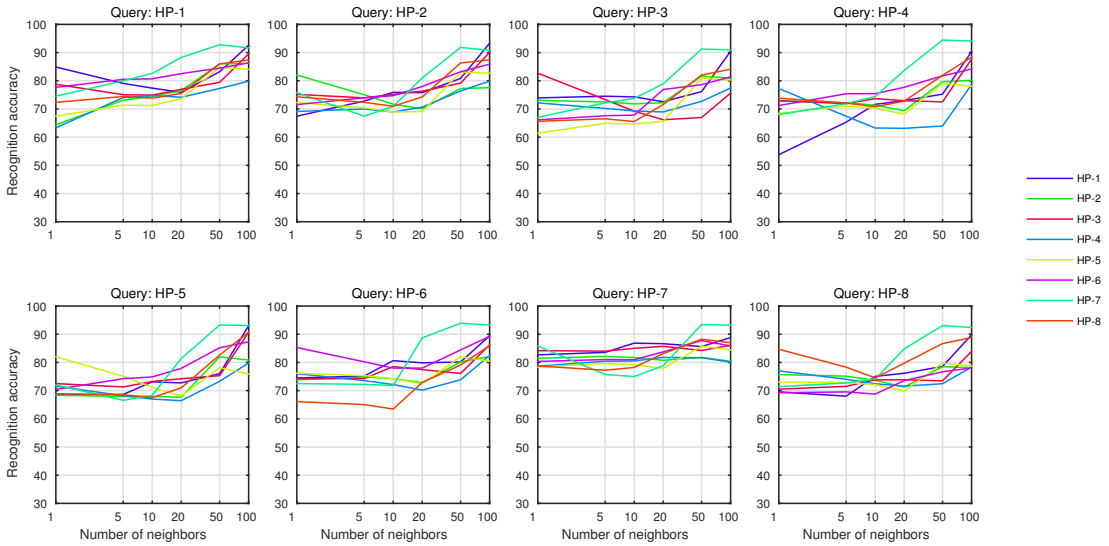



**Figure 4.6**: Full dimensional Fisher Vectors: Pairwise joint metric learning's precision @ $k$ results for across movies retrieval. This figure is best viewed in color.

carried out by linear projection with $W$ matrix. Table 4.13 introduces the mAP results of within and across movies face track retrieval. Compared to the aforementioned baseline results, we also get an improvement in this metric learning by an average equal to 32.9%. The gap between mAP values across movies also similar to the ones obtained by joint metric learning with small standard deviation.

Figure 4.7 refers to the precision @$k$ results. As previously explained, the lines represents the movies and the title indicate the query's' movie. We also can notice that the gap between query and database movies in face track classification is minimized and the precision results are improved compared to the baseline results. The average difference of precision @k for different k values between low-rank Mahalanobis metric learning and baseline is shown in table 4.14. The table shows that there is a progress in all precision of different k values.

**Table 4.12**: Full dimensional Fisher vectors: The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values.

| k values | @1 | @2 | @10 | @20 | @50 | @100 |
|---|---|---|---|---|---|---|
| Improvement Percentage | 14.5 | 19.5 | 22.9 | 26.6 | 34.7 | 40.9 |

**Table 4.13**: Pairwise low rank Mahalanobis metric's cross movies FT retrieval: mAP.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 65.9 | 61.7 | 58.5 | 55.9 | 58.2 | 61.1 | 69.6 | 55.3 |
| HP-2 | 60.2 | 62.6 | 53.7 | 48.0 | 55.0 | 56.6 | 63.4 | 49.3 |
| HP-3 | 53.5 | 51.9 | 56.6 | 44.3 | 46.5 | 48.4 | 57.8 | 42.7 |
| HP-4 | 56.9 | 53.9 | 52.9 | 49.9 | 50.5 | 52.6 | 60.3 | 51.2 |
| HP-5 | 54.6 | 52.4 | 51.9 | 45.8 | 60.2 | 52.1 | 58.3 | 48.6 |
| HP-6 | 56.0 | 54.6 | 53.6 | 49.3 | 54.5 | 59.8 | 59.0 | 45.2 |
| HP-7 | 60.8 | 58.5 | 57.0 | 51.0 | 57.3 | 57.9 | 65.4 | 53.8 |
| HP-8 | 50.6 | 51.7 | 48.3 | 45.8 | 50.1 | 49.9 | 58.1 | 55.0 |

**Diagonal Metric Learning:** last metric learning we use in this scenario is diagonal metric learning. Table 4.15 show the mAP result of across and within movies face track retrieval. Compared to the previous two metric learning, diagonal metric learning has a bad impact on the mAP table for both retrieval tasks (1 and 2) and this metric learning gives slightly worse results according to the baseline results on the reduced dimensional Fisher vectors. The average decline of mAP results in reference to baseline is $-0.7\%$.

Figure 4.8 shows the scores of precision @$k$, as shown on the sub-figures, there is a loss in the values of precision @k scores comparing to the baseline experiment. The gap between query and database movies still exist and even it gets wider. Furthermore, we point out to this decline compared to baseline in table 4.16 which show the average difference of precision @ k between this metric learning and baseline results. There is always loss in the precision for all k values. The reasons behind this trend of the performance is the inability of diagonal metric to learn the similarity features when facial appearance changes exist. It is less complex than the previous metric learning approaches and has a basic objective function with small number of parameters.

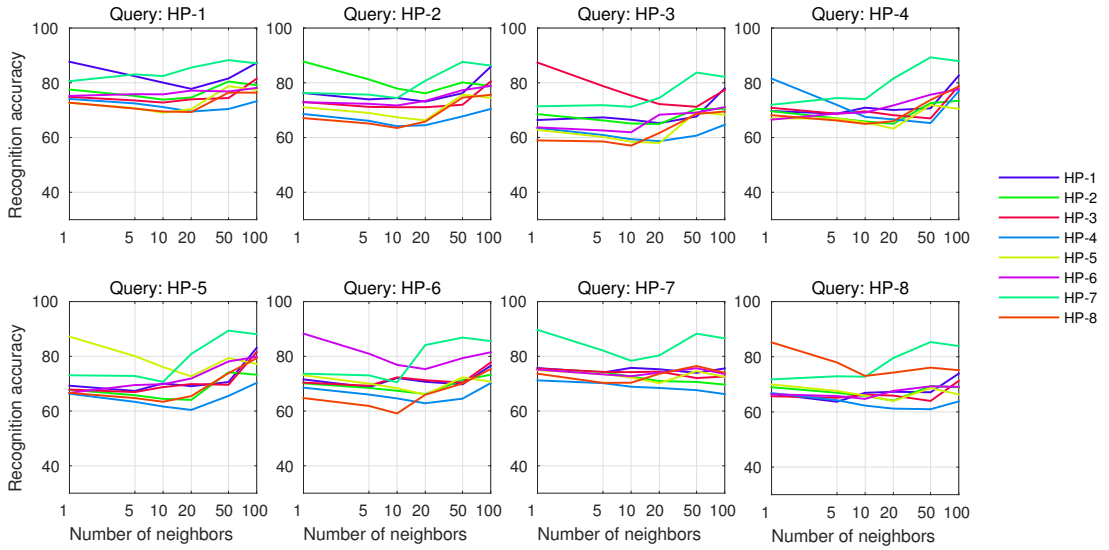



**Figure 4.7**: Pairwise low-Rank Mahalanobis metric learning's precision @ $k$ results for across movies retrieval. This figure is best viewed in color.

**Table 4.14**: The average differences of the precision @k between baseline and low-rank Mahalanobis metric learning that show the improvement in tasks 1 and 2 results for different k values.

| k values | @1 | @2 | @10 | @20 | @50 | @100 |
|---|---|---|---|---|---|---|
| Improvement Percentage | 16.8 | 20.2 | 22.5 | 25.7 | 31.6 | 36.7 |

**Table 4.15**: Pairwise diagonal metric's cross movies FT retrieval: mAP.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 38.5 | 28.4 | 18.1 | 14.5 | 18.0 | 21.6 | 23.7 | 17.8 |
| HP-2 | 27.9 | 23.4 | 22.1 | 15.7 | 15.1 | 21.2 | 25.4 | 12.7 |
| HP-3 | 17.5 | 19.0 | 25.4 | 18.7 | 16.0 | 22.6 | 21.4 | 16.1 |
| HP-4 | 20.6 | 19.1 | 27.0 | 11.3 | 19.5 | 20.8 | 23.0 | 14.0 |
| HP-5 | 22.8 | 16.2 | 20.2 | 17.6 | 28.7 | 26.3 | 27.3 | 20.0 |
| HP-6 | 20.1 | 20.1 | 25.5 | 13.9 | 24.3 | 32.3 | 27.1 | 18.9 |
| HP-7 | 17.6 | 19.6 | 16.0 | 12.8 | 23.1 | 22.4 | 36.2 | 16.7 |
| HP-8 | 16.9 | 16.4 | 19.4 | 11.2 | 19.1 | 21.8 | 22.7 | 28.5 |

**Table 4.16**: The average differences of the precision @k between baseline and diagonal metric learning that show the improvement in tasks 1 and 2 results for different k values.

| k values | @1 | @2 | @10 | @20 | @50 | @100 |
|---|---|---|---|---|---|---|
| Improvement Percentage | -13.8 | -11.2 | -9.6 | -7.6 | -3.9 | -0.5 |

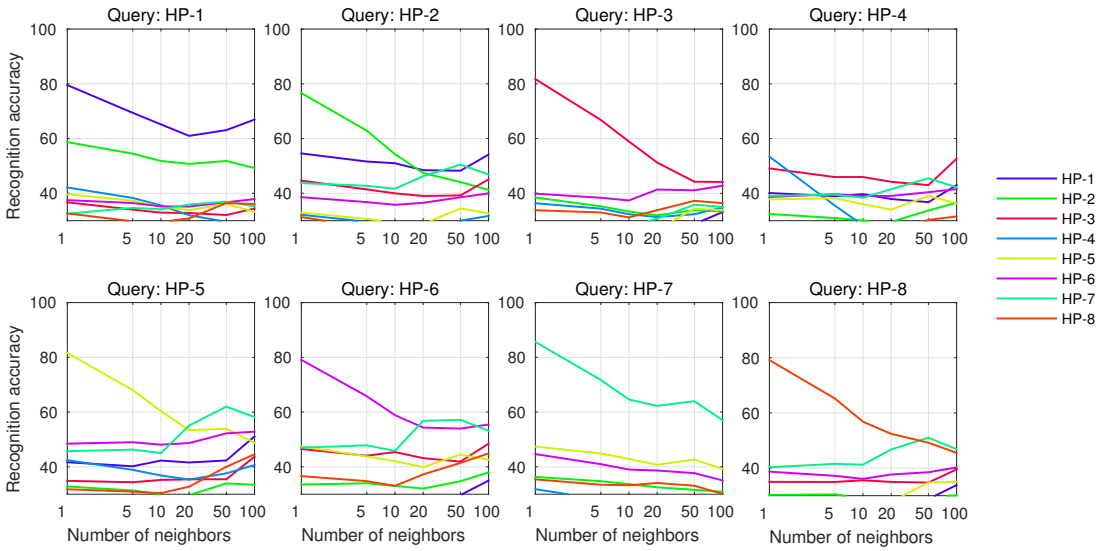
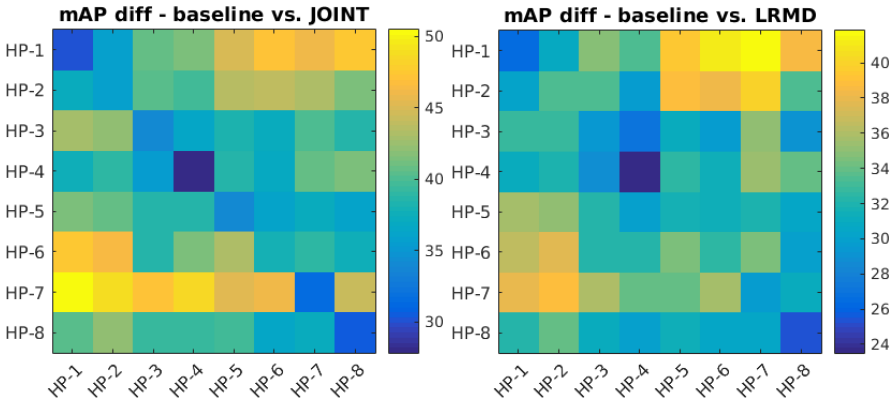



**Figure 4.8**: Pairwise diagonal metric learning's precision @ $k$ results for across movies retrieval. This figure is best viewed in color.

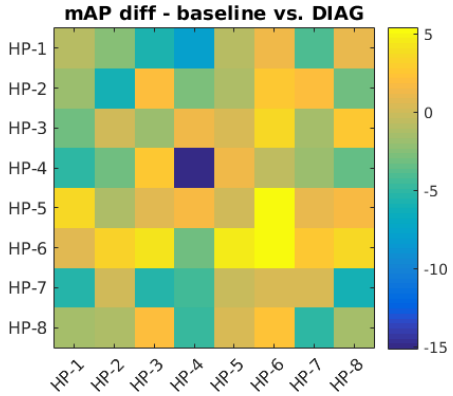
**Confusion Matrices of the Difference Between Baseline and Metric Learning mAPs:**

Figures in 4.9 presents the confusion matrices of the difference between baseline and metric learning mAPs. The biggest improvement is obtained in case of joint metric learning. This due to the fact that joint metric includes low-rank Mahalanobis during learning and the measurement of similarity between pair face tracks arises from the inner product in its distance function. Similarly low-rank Mahalanobis metric gives a great improvement over the baseline mAP. The confusion matrices of these two metric learning show their ability to capture the difference and the similarity between pairwise movies' face tracks which leads to a gain in the performance. However, the diagonal metric learning is less complex compared to low-rank and joint metric learning, that is why its impact on the performance is bad and it leads to a loss in mAP of the retrieval tasks.




(a) Confusion mAP: Baseline - Joint metric

(b) Confusion mAP: Baseline - Low-rank metric




(c) Confusion mAP: Baseline - Diag. metric

**Figure 4.9**: Confusion matrices of difference between baseline and applied pairwise metric learning mAPs.

### 4.3.2 Metric learning on all movies data

In the second type of metric learning that we apply to improve the baseline results of face track retrieval, the training face tracks' data are pooled and combined from all movies to obtain a metric learning model. In this experiment, we fuse 10% of face tracks from each movie selected randomly for training metric learning. The rest 90% face tracks are used for the retrieval across and within movies (tasks 1 and 2). Then using all these face tracks, positive and negative face track pairs are generated regardless to the years gap between movies. As shown in figure 4.1, many characters appear in multiple movies. 11 characters of Harry Potter movies appear in all the Harry Potter 8 movies. This provides huge number of negative and positive face tracks for training metric learning.

The first case in this experiment is the face track retrieval on the test face tracks using the reduced dimensions of Fisher vectors. The results of this case are referred as the baseline results when compared to the retrieval results obtained from metric learning techniques. Table 4.17 gives the mAP results of face track retrieval across and within movies and figure 4.10 presents the precision @$k$ scores for the face track classification across movies. Both mAP and precision @k results prove the decline trend of performance when evaluating across movies.

**Table 4.17**: All movies baseline's cross movies FT retrieval: mAP.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 39.9 | 31.0 | 23.7 | 22.5 | 18.9 | 20.1 | 27.8 | 16.9 |
| HP-2 | 29.9 | 29.7 | 20.2 | 18.3 | 16.3 | 18.4 | 23.4 | 15.9 |
| HP-3 | 20.7 | 19.0 | 28.0 | 17.3 | 15.5 | 18.9 | 22.8 | 13.5 |
| HP-4 | 25.7 | 22.0 | 24.2 | 26.4 | 18.0 | 21.4 | 24.9 | 17.4 |
| HP-5 | 19.0 | 17.3 | 19.5 | 15.8 | 29.1 | 20.9 | 26.3 | 18.2 |
| HP-6 | 19.3 | 16.9 | 21.2 | 17.2 | 19.9 | 27.6 | 24.5 | 15.3 |
| HP-7 | 22.8 | 19.6 | 21.1 | 17.1 | 23.2 | 22.1 | 36.2 | 22.8 |
| HP-8 | 18.5 | 17.7 | 17.2 | 16.0 | 18.8 | 19.3 | 27.5 | 30.1 |

**Joint Metric Learning:** In this experiment, after combining the training data from all movies the projection matrices $(W,V)$ with a dimension equal to $1000 \times 128$ are trained by the negative and positive face track pairs. Then we use these projection

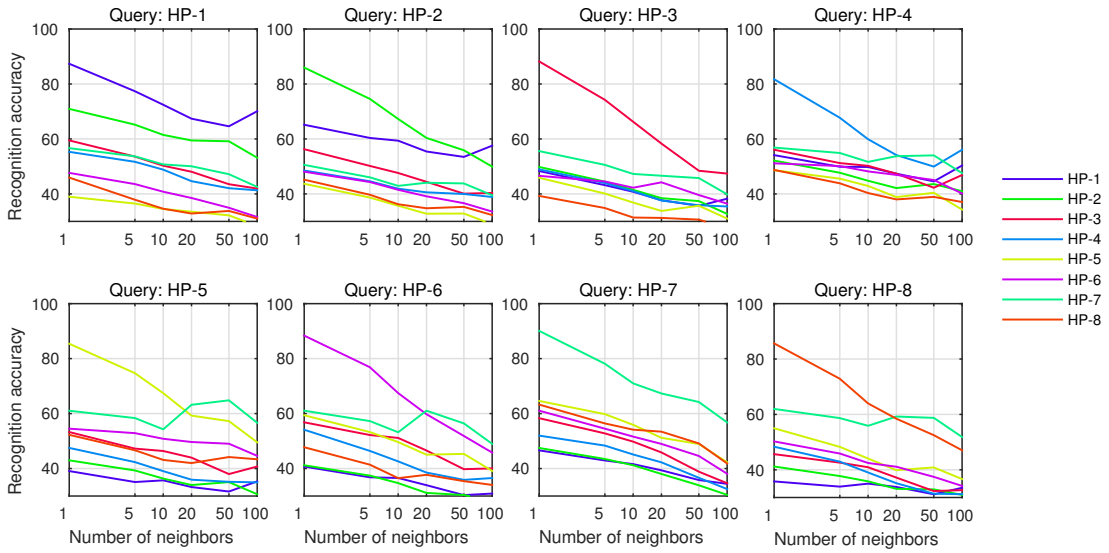



**Figure 4.10**: FT from all movies: Baseline's precision @ $k$ results for across movies retrieval. This figure is best viewed in color.

matrices for the two retrieval tasks, 1 and 2. Table 4.18 gives the mAP results of face track retrieval across and within movies. Compared to the baseline table in 4.17, there is an improvement in all mAP results by average equal to 42.7%. For example, for the mAP of across movies retrieval between HP-1 and HP-8, we have an improvement over the baseline mAP by an average equal to 49.7%, for example the improvement in mAP between HP-1 and HP-2 is 30.4%. In addition, the standard deviation of the across and within movies retrieval mAP is lower than the one in baseline mAP. This show clearly the effect of combining face tracks from all movies since the results are also better that the pairwise metric learning introduced in table 4.9 by 2.7%.

**Table 4.18**: All movies joint metric's cross movies FT retrieval: mAP.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 66.7 | 67.9 | 65.9 | 66.7 | 65.2 | 72.9 | 77.5 | 66.6 |
| HP-2 | 68.4 | 62.3 | 63.6 | 61.0 | 63.3 | 66.5 | 71.1 | 63.6 |
| HP-3 | 65.5 | 64.5 | 57.8 | 60.0 | 57.1 | 62.1 | 68.5 | 58.4 |
| HP-4 | 65.4 | 64.5 | 62.3 | 51.0 | 58.8 | 63.2 | 68.3 | 62.9 |
| HP-5 | 62.9 | 61.4 | 58.0 | 55.4 | 57.1 | 59.1 | 63.4 | 56.9 |
| HP-6 | 70.7 | 68.3 | 64.8 | 63.7 | 65.1 | 61.7 | 67.0 | 56.3 |
| HP-7 | 77.7 | 74.3 | 72.3 | 68.8 | 69.7 | 73.6 | 65.7 | 67.8 |
| HP-8 | 62.1 | 65.9 | 61.6 | 60.7 | 61.2 | 62.4 | 65.6 | 60.6 |

Furthermore, figure 4.11 shows the precision @$k$ scores for face track classification. Notice that the lines are much closer to each other regardless to the query movie. This

is due to bringing all face tracks to one domain and train the joint metric learning on these data. In addition, not only the gap between lines in sub-figures is reduced, but also the precision scores of different $k$ values as shown in table 4.19 are improved even for large $k$ values. For $k = 50, 100$ the results get much better due the fact that, not all characters have that number of face tracks in each movie. Only major characters such as Harry Potter, Hermione and Ron Weasley have that number of face tracks. When combining the data for the training goal, large number of these face tracks of training belong to these characters too. This is why most of face tracks are correctly classified even when dealing with high value of $k$.

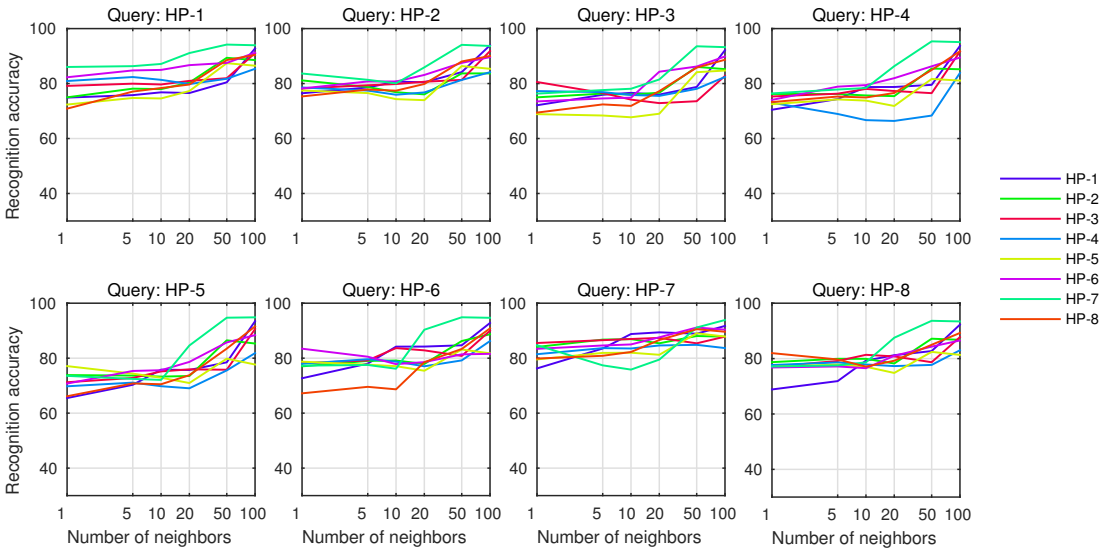



**Figure 4.11**: FT from all movies: Joint's precision @ $k$ results for across movies retrieval. This figure is best viewed in color.

**Table 4.19**: Metric learning from from all movies: The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values.

| k values | @1 | @2 | @10 | @20 | @50 | @100 |
|---|---|---|---|---|---|---|
| Improvement Percentage | 20.5 | 27.1 | 30.7 | 34.7 | 42.2 | 48.7 |

**Joint Metric Learning on Full Dimensional Fisher Vectors:** We also give the results of retrieval tasks in this scenarios using the full dimensions of face track descriptors. Table 4.20 show the mAP scores of two retrieval tasks. We observe that wen we apply the metric learning on reduced dimensional face tracks descriptors, we get better results. As previously stated, PCA reduce the noise and the redundancy in the data such that metric learning is more capable to highlight the similarity features in better way.

Figure 4.12 also shows the precision @k results of face track classification. We also notice the impact of PCA here since the gap between lines is higher and the precision @k values are also decreased. Table 4.21 proves that since the average difference of precision @k between baseline and joint metric learning is less than the average difference when we use the reduced dimensional face track descriptors.

**Table 4.20**: Full dimensional Fisher vectors - All movies joint metric's cross movies FT retrieval: mAP.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 61.8 | 63.8 | 63.3 | 60.4 | 60.5 | 67.3 | 76.2 | 63.9 |
| HP-2 | 64.9 | 59.3 | 62.1 | 57.4 | 60.6 | 62.5 | 70.3 | 61.7 |
| HP-3 | 61.8 | 62.0 | 57.1 | 56.2 | 54.6 | 59.1 | 68.0 | 56.3 |
| HP-4 | 61.5 | 61.5 | 59.9 | 46.9 | 55.5 | 59.8 | 67.4 | 60.8 |
| HP-5 | 59.7 | 59.7 | 56.5 | 52.0 | 54.0 | 56.4 | 62.4 | 55.6 |
| HP-6 | 66.7 | 64.5 | 62.7 | 59.8 | 61.8 | 58.5 | 65.7 | 53.2 |
| HP-7 | 73.7 | 72.6 | 70.8 | 64.7 | 66.8 | 69.7 | 64.4 | 66.4 |
| HP-8 | 59.1 | 64.1 | 60.1 | 57.0 | 57.7 | 59.2 | 64.8 | 57.4 |

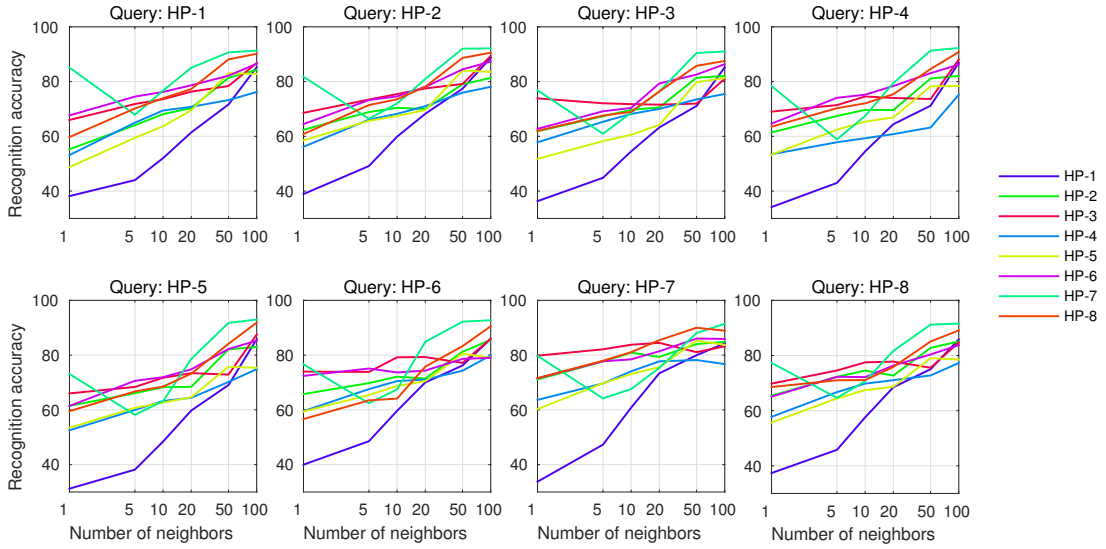



**Figure 4.12**: FT from all movies: Joint's precision @ $k$ results for across movies retrieval. This figure is best viewed in color.

**Low-rank Mahalanobis Metric Learning:** In this case, a projection matrix $W$ with a dimension = 128 is learned on the combined 10% face tracks using the generated positive and negative pairs. As aforementioned joint metric learning, low-rank Mahalanobis metric learning achieves good results comparing the baseline mAP results in both retrieval tasks, across and within movies face track retrieval. Table 4.22 presents the mAP scores which shows that the retrieval across movies are increased

**Table 4.21**: Metric learning from from all movies on full dimensional Fisher vectors: The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values.

| k values | @1 | @2 | @10 | @20 | @50 | @100 |
|---|---|---|---|---|---|---|
| Improvement Percentage | 2.1 | 11.3 | 18.3 | 24.9 | 33.8 | 40.2 |

with an average difference equal to 35.5%, and the gap that appears in table 4.17 between across movies mAPs is also reduced and the standard deviation as well.

**Table 4.22**: All movies low-rank Mahalanobis metric's cross movies FT retrieval: mAP.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 63.9 | 61.4 | 60.8 | 56.0 | 58.9 | 63.7 | 72.7 | 58.6 |
| HP-2 | 61.0 | 58.1 | 57.7 | 50.0 | 55.9 | 58.3 | 67.2 | 54.8 |
| HP-3 | 56.9 | 55.0 | 56.3 | 49.2 | 51.5 | 53.4 | 63.3 | 49.8 |
| HP-4 | 57.5 | 56.0 | 57.5 | 45.9 | 51.9 | 54.4 | 63.0 | 53.5 |
| HP-5 | 57.0 | 54.8 | 55.7 | 47.7 | 56.1 | 53.8 | 61.4 | 50.8 |
| HP-6 | 60.5 | 58.1 | 59.2 | 52.2 | 57.3 | 58.2 | 64.3 | 48.0 |
| HP-7 | 65.7 | 62.2 | 63.9 | 54.2 | 60.6 | 62.3 | 65.7 | 57.4 |
| HP-8 | 54.5 | 56.2 | 56.1 | 48.4 | 52.9 | 54.3 | 61.7 | 55.0 |

In addition, figure 4.13 presents the precision @$k$ scores for the face classification task. This metric learning bring the lines of query and databases movies to the same range and also improve the precision scores in different $k$ values. Even though the improvement is not as much as the joint metric learning scores shown in figure 4.11, it is better that the baseline results (4.10) in term of the precision scores and the gap between lines of queries and databases. The precision scores at all k values is also increased as we display in table 4.23

**Table 4.23**: Metric learning from from all movies: The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values.

| k values | @1 | @2 | @10 | @20 | @50 | @100 |
|---|---|---|---|---|---|---|
| Improvement Percentage | 19.5 | 23.1 | 25.4 | 28.7 | 34.1 | 39.0 |

**Diagonal Metric Learning:** in the last case, we use diagonal metric learning to obtain a metric model from the combined 10% face tracks. Table 4.24 shows the mAP scores

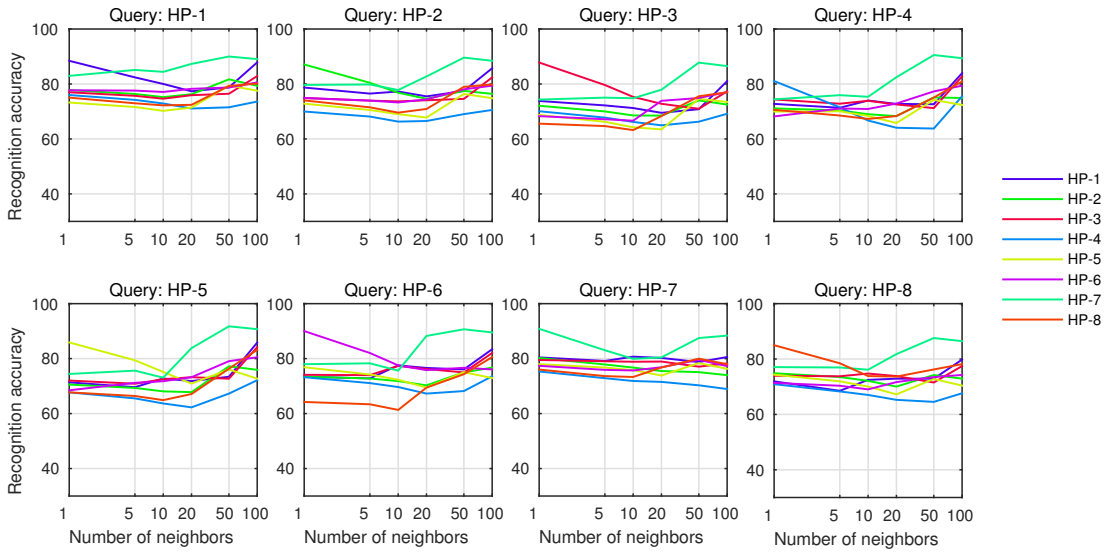



**Figure 4.13**: FT from all movies: Low-rank Mahal.'s precision @ $k$ results for across movies retrieval. This figure is best viewed in color.

of face track retrieval across and within movies. Compared to the baseline results in table 4.17, there is a loss in the performance by an average equal to $-2.5\%$ in this metric learning. This metric learning tries to learn a weight for each dimension of Fisher vectors and uses the conventional objective function of linear SVM. It turns out that, bringing all face tracks from different domains which are the eight movies introduce a challenging data such that diagonal metric learning can not capture that similarity of faces in different movies. Note that, not only the values of mAP scores are decreased, but also the behavior of mAP is identical to the one in the baseline where the diagonal elements are the highest which represent the within movies face track retrieval. As we go farther from the diagonal elements when the retrieval done across movies the performance decreases.

This fact shows the effect of using joint and low-rank Mahalanobis metric learning in this experiment, these two metric learning learned the similarity features of faces in different years since the training data is gathered from the eight movies which are released in a period of ten years. The training data contains a facial changes of the characters across age.

Finally figure 4.14 introduces the precision @$k$ scores of face classification across movies. It shows the the overall performance is degraded and the precision values are decreased compared the baseline figure 4.10. Table 4.25 also displays the decline in the average difference of precision between baseline and diagonal metric results. This is

**Table 4.24**: All movies diagonal metric's cross movies FT retrieval: mAP.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 27.6 | 20.6 | 20.3 | 16.2 | 18.0 | 18.3 | 28.8 | 15.5 |
| HP-2 | 21.6 | 20.8 | 19.2 | 14.7 | 17.1 | 17.1 | 24.1 | 14.7 |
| HP-3 | 20.1 | 17.2 | 21.4 | 13.8 | 15.3 | 16.9 | 23.9 | 12.9 |
| HP-4 | 22.1 | 19.4 | 20.9 | 17.0 | 19.2 | 19.5 | 27.4 | 16.7 |
| HP-5 | 20.7 | 17.9 | 19.2 | 15.6 | 20.9 | 17.9 | 23.2 | 14.6 |
| HP-6 | 19.2 | 16.4 | 19.7 | 14.5 | 18.3 | 22.5 | 23.7 | 12.5 |
| HP-7 | 25.5 | 20.8 | 22.7 | 17.1 | 19.8 | 20.5 | 28.6 | 17.5 |
| HP-8 | 17.2 | 16.6 | 17.4 | 13.9 | 15.2 | 15.7 | 21.6 | 20.6 |

due to the aforementioned reasons which doubt the ability of diagonal metric learning to learn a model that find the similarity of face features in different domains.

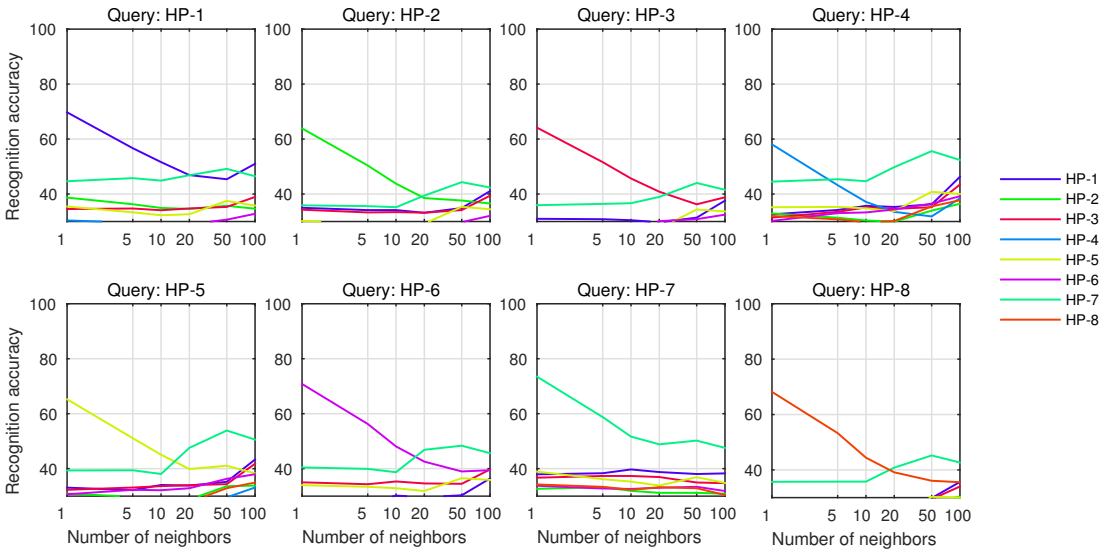
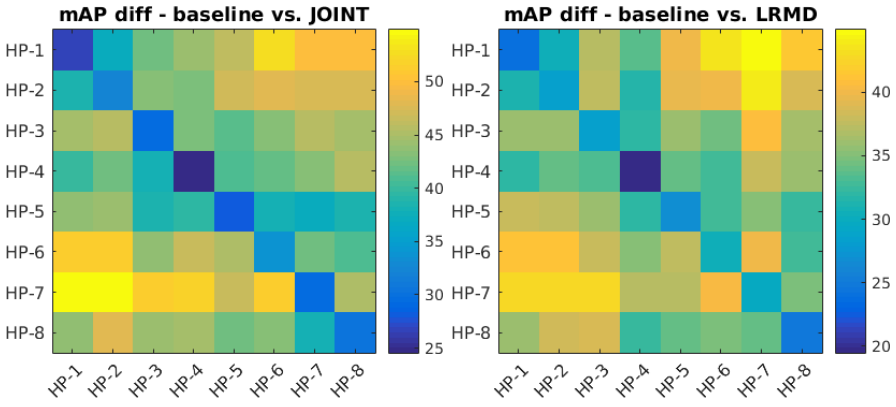



**Figure 4.14**: FT from all movies: Diagonal metric's precision @ $k$ results for across movies retrieval. This figure is best viewed in color.

Figures in 4.15 show the confusion matrices of the difference between baseline and applied metric learning mAPs. Each figure has a color bar next which indicates the improvement values. It turns out that joint and low-rank Mahalanobis metric techniques have an improvement over the baseline mAP, while diagonal metric results in reduction in the performance as previously mentioned.

**Table 4.25**: Metric learning from from all movies: The average differences of the precision @k between baseline and joint metric learning that show the improvement in tasks 1 and 2 results for different k values.

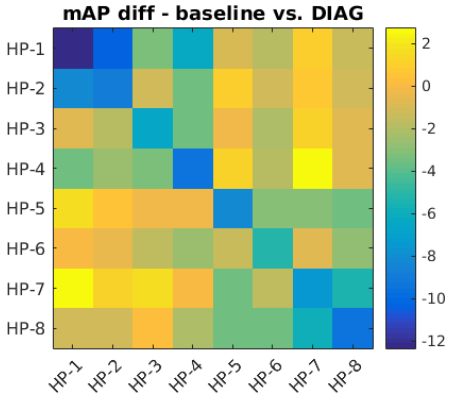
| k values | @1 | @2 | @10 | @20 | @50 | @100 |
|---|---|---|---|---|---|---|
| Improvement Percentage | -19.8 | -16.4 | -14.3 | -11.9 | -7.6 | -3.6 |




(a) Confusion mAP: Baseline - Joint metric




(b) Confusion mAP: Baseline - LRMD metric




(c) Confusion mAP: Baseline - Diag. metric

**Figure 4.15**: Difference between baseline mAP and applied data from all movies metric learning mAPs.

## 4.4 Face Track Recognition

In this section, we explain the supervised learning applied on Accio dataset for face track recognition using linear SVMs. Prior to face track recognition experiment, a character of Accio dataset in any movie is included into the evaluation if it appears at least in three of Harry Potter movies and has at least a total number of 50 face tracks across movies. we set these criteria in order to obtain multi-class linear SVM for training for each character in the dataset since we want to have balanced number of face tracks for all characters included in our evaluation. Figure 4.16 shows the montage image of sample characters that satisfy these criteria.

For the evaluation we randomly select 5 fold cross validation sets, in each fold, one set of tracks is used for training linear SVMs and the rest are for the test. A training fold contains 20% of face tracks and the test set contains the rest of face tracks Similarly to face track retrieval, we have two task for the evaluation:

(i) *Within* movies: In this setting, we evaluate movies separately at a time. The evaluation is conducted only on named face tracks with specific criteria. This setting evaluates the recognition performance subject to typical face challenges such as variant pose and illumination.

(ii) *across* movies: In this setting we analyze the impact of age progression among the actors evaluating face track recognition across movies. We also only consider named face tracks across movies for both training and testing. The evaluation of across movies is achieved in pairwise different movies (for example HP-1 and HP-2 ).

**Figure 4.16**: Sample images for characters that satisfy that SVM training criteria.

Following the fulfillment of the criteria, a linear SVM model is obtained for each character in a movie. For example, there are 23 characters as shown in table 4.3, this model then is used for evaluation for both tasks: within the same movie and across the other movies. The evaluation is done for the 5 folds in cross validation sets and the performance results are the average of the 5 folds. Table 4.26 shows the overall accuracies of cross and within movies face recognition. Each accuracy shown in the table, it is calculated by combining all scores from SVM models of a train movie and then compare the highest score with the test label of test movie, if it is the same, then the we decide that the character is correctly classified, and false otherwise.

The table 4.26 indicates that the best results are obtained when the test is applied within the same movie where only standard challenges such as illumination and pose exist without age variation. However, as we go farther from the diagonal elements, to the right or to the left, we can see how the accuracy values decrease as the time gap between the SVM models and test data of movies increases.

**Table 4.26**: SVM accuracy: overall.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 91.0 | 82.1 | 71.0 | 67.7 | 56.7 | 59.7 | 56.7 | 60.1 |
| HP-2 | 75.7 | 90.4 | 66.2 | 60.0 | 61.6 | 61.2 | 52.1 | 59.2 |
| HP-3 | 64.2 | 64.9 | 85.7 | 58.9 | 60.6 | 55.9 | 53.8 | 56.9 |
| HP-4 | 64.6 | 63.5 | 60.3 | 87.2 | 68.6 | 63.6 | 56.7 | 63.8 |
| HP-5 | 50.8 | 56.9 | 55.2 | 59.3 | 87.0 | 57.4 | 53.1 | 58.4 |
| HP-6 | 63.7 | 67.8 | 64.8 | 65.2 | 73.2 | 86.9 | 59.6 | 68.9 |
| HP-7 | 65.3 | 68.6 | 68.4 | 68.8 | 80.2 | 73.0 | 89.3 | 78.0 |
| HP-8 | 51.8 | 56.1 | 53.1 | 64.0 | 70.4 | 62.1 | 62.2 | 87.6 |

Tables 4.27, 4.28, and 4.29 presents the F1 scores of face recognition of the three major characters in Accio dataset: Harry Potter, Hermione, and Ron Weasley. F1 score consider the precision and recall, where precision is the number of correct positive face tracks divided all positive face tracks and recall is the number of correct positive face tracks divided by the positive tracks that should be identified. F1 score is the weighted average of precision and recall as given in equation 4.1:

$$F_1 = 2.\frac{\text{recall . precision}}{\text{recall + precision}} \quad \textbf{(4.1)}$$

**Table 4.27**: F1 score of face recognition for Harry Potter character in Accio dataset.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 95.8 | 93.1 | 87.7 | 85.5 | 67.2 | 77.2 | 76.8 | 79.3 |
| HP-2 | 88.2 | 95.9 | 85.0 | 81.2 | 73.3 | 78.0 | 74.3 | 78.3 |
| HP-3 | 85.0 | 87.6 | 93.6 | 82.6 | 84.9 | 81.8 | 80.7 | 80.4 |
| HP-4 | 90.3 | 90.9 | 90.3 | 94.8 | 88.6 | 86.6 | 85.4 | 87.3 |
| HP-5 | 79.5 | 83.5 | 84.4 | 82.0 | 94.0 | 83.0 | 80.8 | 84.0 |
| HP-6 | 86.4 | 88.1 | 89.7 | 85.0 | 91.2 | 94.2 | 87.0 | 87.4 |
| HP-7 | 84.3 | 84.4 | 87.0 | 83.7 | 88.5 | 86.9 | 94.5 | 87.9 |
| HP-8 | 78.4 | 78.7 | 81.4 | 80.0 | 86.7 | 81.0 | 83.9 | 93.2 |

**Table 4.28**: F1 score of face recognition for Hermione character in Accio dataset.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 91.6 | 86.2 | 75.0 | 53.2 | 67.9 | 50.5 | 44.1 | 47.5 |
| HP-2 | 76.7 | 91.4 | 66.1 | 45.7 | 63.6 | 53.4 | 32.7 | 48.8 |
| HP-3 | 74.1 | 78.6 | 87.3 | 57.7 | 74.8 | 60.0 | 57.4 | 66.0 |
| HP-4 | 61.4 | 60.1 | 47.9 | 83.5 | 76.1 | 67.6 | 50.2 | 61.3 |
| HP-5 | 56.8 | 62.8 | 54.2 | 61.9 | 85.4 | 63.5 | 45.8 | 57.1 |
| HP-6 | 58.2 | 66.2 | 61.3 | 65.3 | 70.6 | 86.9 | 52.4 | 62.2 |
| HP-7 | 73.6 | 74.5 | 76.3 | 80.3 | 85.6 | 81.5 | 88.7 | 82.5 |
| HP-8 | 64.2 | 71.8 | 64.6 | 70.5 | 81.9 | 69.0 | 62.0 | 86.8 |

**Table 4.29**: F1 score of face recognition for Ron Weasley character in Accio dataset.

| | HP-1 | HP-2 | HP-3 | HP-4 | HP-5 | HP-6 | HP-7 | HP-8 |
|---|---|---|---|---|---|---|---|---|
| HP-1 | 91.6 | 79.1 | 69.1 | 71.3 | 56.5 | 58.6 | 63.7 | 62.1 |
| HP-2 | 78.5 | 91.8 | 73.5 | 47.6 | 69.7 | 71.1 | 59.0 | 66.6 |
| HP-3 | 60.0 | 61.5 | 84.8 | 42.7 | 58.7 | 60.9 | 52.9 | 56.3 |
| HP-4 | 66.0 | 49.1 | 38.9 | 91.6 | 61.2 | 63.9 | 60.3 | 59.0 |
| HP-5 | 48.7 | 63.9 | 62.7 | 51.5 | 90.5 | 67.8 | 68.6 | 63.7 |
| HP-6 | 47.9 | 76.2 | 68.1 | 38.0 | 83.1 | 89.3 | 68.6 | 71.6 |
| HP-7 | 58.1 | 67.0 | 66.9 | 61.7 | 87.4 | 82.6 | 92.6 | 86.1 |
| HP-8 | 43.2 | 57.4 | 56.5 | 58.6 | 81.3 | 77.0 | 77.0 | 90.1 |

# 5. FACE VERIFICATION ANALYSIS USING FISHER VECTORS

The aim of this chapter is to evaluate the performance of Fisher vector on face verification using different dataset and features rather than LFW dataset and SIFT features. We apply evaluation on the fourth experiment of Face Recognition Grand Challenge(FRGC) dataset. We first explain the dataset and its performance evaluation metric, and then we discuss the usage of Gabor filters. Finally we provide all the details of experiments results with variant Fisher vector settings.

## 5.1 Face Recognition Grand Challenge(FRGC) Data Set

FRGC is contains three sub sets of high resolution still images taken under controlled and uncontrolled conditions. Controlled images were taken in a studio settings while the uncontrolled images were taken with variance illumination conditions, with two expression: smiling and neutral. In addition, FRGC contains several experiment's settings. In this work, we use the fourth experiment which measures the progress on recognition from uncontrolled frontal still images[4].

**FRGC's Fourth Experiment:** The validation sets of FRGC's fourth experiment involves two sets, query and target sets. Query set consists of 8,014 uncontrolled single images while the target set consists of 16,028 single controlled images.The two sets contain images from 466 identities. The training set of FRGC consists of 12,776 images from 222 subjects, with 6,388 controlled and 6,388 uncontrolled images. This is the most challenging experiment due to the uncontrolled settings which have variant illumination, focus change and partial occlusion.

**Performance Evaluation** The reference point to measure the performance of a method on FRGC is the verification rate (VR) at 0.1% false acceptance rate. In addition there are three Receiving Operator Characteristic (ROC) curves corresponding to three different timing gap. ROCI corresponds to the images taken within a semester, ROCII within a year and ROCIII corresponds to images captured between two semester.

**Figure 5.1**: Still images taken in one session under controlled conditions (Studio Settings) [4].

**Figure 5.2**: Two uncontrolled still images[4].

## 5.2 Experiments: FRGC's Fourth Experiment

### 5.2.1 SIFT features

First of all the pipeline of Fisher vector faces in the wild as suggested in[3] is applied on FRGC's fourth experiment using SIFT features and as explained in chapter 3.3.1. The face verification starts with dense sift extraction, reducing the dimensionality of face descriptor by PCA, training GMM and then encoding Fisher vectors. It is important to mention that the reported results of LFW sets are re-obtained in during this work as presented in the following table 5.1.

**Table 5.1**: Evaluation results of LFW splits in term of ROC = ( 1 - EER).

| Feats. | Aug. | GMM Size | PCA Dim | Applied Metric Learning | ROC% |
|---|---|---|---|---|---|
| SIFT | ✓ | 512 | 64 | Low Rank Mahalanobis Metric | 91.66 |
| SIFT | ✓ | 512 | 64 | Low Rank Joint Metric | 91.4 |
| SIFT | ✓ | 512 | 64 | Diagonal Metric | 90.9 |

For the evaluation of FRGC's fourth experiemnt, the only difference is the size of GMM which is set to 128. This gives the best results and the computation of Fisher vectors become faster since their dimensions depends on GMM size (K). The resulted

Fisher vectors dimensions $= 2PK = 2 * 66 * 128 = 16896$. The Root dense-SIFT settings are as the following: patch size is $24 * 24$, scales = 5 with a scale factor equal to $\sqrt{2}$. Since the face images are aligned and set to $128 * 160$ the SIFT features has $12828K$ dimensions, which they are reduced by PCA into $6428K$ dimensions and then augmented with the spatial information. The results are expressed in term of AUC and VR@0.1% FAR and shown in table 5.2. In addition we applied the large-margin dimensionality reduction methods in order to make Fisher vectors compressed and more discriminative along with diagonal and joint metric learning.

**Note:** It is important to note that zero-normalization is applied on the verification scores by subtracting the mean of score matrix and dividing by the standard deviation. This normalization always helps to improve the face verification performance. In the following tables, we refer to zero-normalization as Z-norm in the first row and we present the AUC and VR@FAR=0.1% results prior and after applying this normalization to display its impact on the performance.

In addition, the results displayed in the following tables for all evaluations are the average of the three ROC curves of the FRGC's fourth experiment.

**Table 5.2**: Using SIFT features: AUC and VR@0.1 FAR% of FRGC's fourth experiment.

| Feats. | Aug. | GMM Size | PCA Dim | Metric Learning | AUC | VR @FAR=0.1% | Z-norm AUC | Z-norm VR@FAR=0.1% |
|---|---|---|---|---|---|---|---|---|
| SIFT | ✓ | 512 | 64 | **Mah. Metric** | 99.26 | 51.0 | **99.4** | **63.9** |
| SIFT | ✓ | 512 | 64 | Joint Metric | 99.3 | 60.1 | 99.2 | 55.1 |
| SIFT | ✓ | 512 | 64 | Diag. Metric | 98.4 | 40.8 | 98.6 | 48.8 |

### 5.3 Fisher Vectors Using Gabor Filters

In addition to the aforementioned Fisher vectors evaluation, several experiments were done in order to assess the performance of Fisher vectors using Gabor filters as a local descriptors. Local blocks of Gabor filters are used for Fisher vector encoding rather than SIFT. Gabor filters are chosen as local features since they are one of the most successful face representations due to their biological relevance to human . They decompose the given face images into several scales and orientations. For a given face images, first it is aligned as shown in figure 5.3. FRGC dataset provide the coordinates

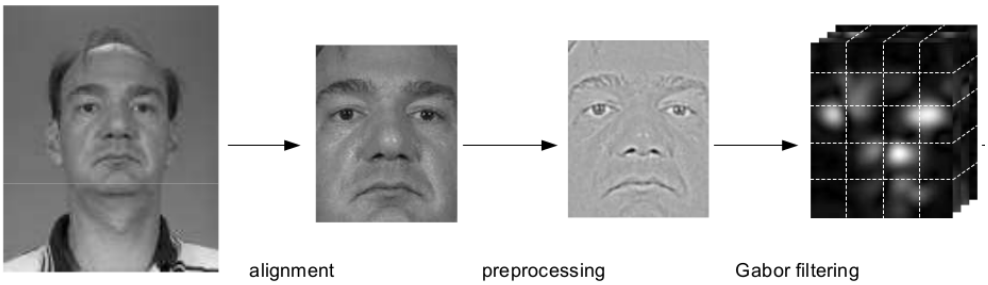



**Figure 5.3**: Pipeline of Gabor filters representations for face image [5].

points of eyes centers and mouth corners which are used for the alignment by similarity transformation. The aligned face image also is resized into $128x160$ pixel resolution. Then a preprocessing is applied which consists of a sequence of the following steps: Gamma correction, Difference of Gaussian (DoG) and highlight suppression. The aligned and preprocessed face image is convolved with a 2-D Gabor wavelet filter bank to obtain a set of filter response with different scale and orientations. This process's steps are illustrated in figure 5.3.

The resulting GMI is divided into 20 non-overlapped blocks with a size $32 \times 32$. Each block is down-sampled with scale factor equal to 0.5 which makes each block size equal to $16 \times 16$. This leads into 256 dimensional vector for each block. Since we have 20 blocks for each image and 40 Gabor responses, the local descriptor of Gabor filters for face image is equal to $256 \times 800$ dimensions.

**Feature Normalization:** Following the extraction of Gabor features, the local descriptors are normalized by taking the square root and dividing them by features norm.

### 5.3.1 Spatial, scale and orientations augmentations:

Since Fisher vector does not capture the local distribution and structure of features, the spatial information and the corresponding scale and orientation are added to the local descriptors. The resulting local descriptor can be formulated as the following: $[G_{x,y}; x; y; S; O]$, where $G_{x,y}$ are the PCA-Gabor filters features of a block centered at $(x,y)$ which has $(S,O)$ scale and orientations.

## 5.4 Experiments and Results

In this section, we carry out several experiments with variant parameters for Fisher vectors in order to analyze the performance of the method under different settings. In the following experiments, one parameter is changed during the experiment while the rest of parameters and settings are kept fixed in order to check the performance of the method related to the changed parameters.

### 5.4.1 Effect of face alignment

In this experiment the aim is to measure the effect of face alignment on the performance of Fisher vectors, figure 5.4, shows two completely different aligned face, where face a is aligned fitly and face b is aligned with a much freedom of background. By fixing the the GMM size to 512, PCA dimensions to 32 augmentation, and using two different face alignments, table 5.3 presents the results of this experiment. According to the results in the table, the difference of the two alignment is not much due to the fact that Fisher vectors automatically captures the important facial features without explicitly trained to do so.

Following this experiment, we used the alignment with **a** background in the rest of the work.

(a) Alignment with Background (b) Fit alignment

**Figure 5.4**: Two different aligned faces.

**Table 5.3**: Effect of face alignment: AUC and VR@0.1% FAR of FRGC's fourth experiment.

| Feats | Align. | Aug. | GMM Size | PCA Dim | Metric Learning | AUC | VR@ FAR=0.1% | Z-norm AUC | Z-norm VR @FAR=0.1% |
|---|---|---|---|---|---|---|---|---|---|
| Gabor | a | ✓ | 512 | 32 | **Mah. Metric** | 98.90 | 42.87 | **99.1** | 57.36 |
| Gabor | a | ✓ | 512 | 32 | Joint Metric | 98.34 | 38.38 | 98.38 | 38.60 |
| Gabor | a | ✓ | 512 | 32 | Diag. Metric | 97.34 | 41.90 | 97.62 | 46.95 |
| Gabor | b | ✓ | 512 | 64 | **Mah. Metric** | 98.8 | 36.9 | 99.0 | **58.3** |
| Gabor | b | ✓ | 512 | 64 | Joint Metric | 98.0 | 37.2 | 98.1 | 38.9 |
| Gabor | b | ✓ | 512 | 64 | Diag. Metric | 96.1 | 37.0 | 97.0 | 41.6 |

### 5.4.2 Effect of PCA dimension size

Preceding training GMM and encoding Fisher vectors on 256$x$800 dimensional Gabor filters local descriptors, we perform PCA for dimensionality reduction to make Fisher vectors applicable for evaluation of FRGC and learning metric models. In this experiment we applied PCA to Gabor local descriptors to decrease their dimensionality from 256 to 32 and 64. We tried to use two different PCA dimensions to check how would this affect the performance of Fisher vectors. Table 5.4 shows the results of using two different dimensions of PCA, while the rest of Fisher vector settings are fixed in both cases. It turns out that lower dimensions are more decorrelated and clear of data redundancy such that the performance of 32 dimensional Fisher vectors is better.

**Table 5.4**: Effect of PCA Dimensions AUC and VR@0.1% FAR of FRGC's fourth experiment.

| Feats. | Aug. | GMM Size | PCA Dim | Metric Learning | AUC | VR @FAR=0.1% | Z-norm AUC | Z-norm VR @FAR=0.1% |
|---|---|---|---|---|---|---|---|---|
| Gabor | ✓ | 512 | 32 | **Mah. Metric** | 98.9 | 42.9 | **99.1** | **57.4** |
| Gabor | ✓ | 512 | 32 | Joint Metric | 98.3 | 38.4 | 98.4 | 38.6 |
| Gabor | ✓ | 512 | 32 | Diag. Metric | 97.3 | 41.9 | 97.6 | 46.9 |
| Gabor | ✓ | 512 | 64 | Mah. Metric | 98.7 | 40.2 | 99.0 | 54.8 |
| Gabor | ✓ | 512 | 64 | Joint Metric | 98.2 | 32.8 | 98.2 | 32.8 |
| Gabor | ✓ | 512 | 64 | Diag. Metric | 97.1 | 40.9 | 97.5 | 46.1 |

### 5.4.3 Effect of spatial, scale and orientation augmentation

Fisher vectors are efficient encoding method however since the features are extracted from face locally, it does not capture the distribution of features in spatial domain. Furthermore, Gabor responses corresponding to different scales and orientations.

Spatial, scales and orientations information should explicitly incorporated into Fisher vectors encoding as suggested in the literature. In this experiment we show how clearly the augmentation of these information affects the performance of Fisher vectors on FRGC's fourth experiment. The augmentation provides the visual information of Gabor responses to GMM which makes the encoding of Fisher vectors more effective. The augmentation is applied as explained in 5.3.1 section. Table 5.5 shows the huge difference of performance when augmenting the spatial, scale, and orientation information to Gabor-PCA features. We can see clearly that the augmentation of these information improve the performance of Fisher vectors. In this experiment the rest settings of Fisher vectors such as GMM size, PCA dimensions, face alignment are fixed in both cases: with augmentation and without augmentation.

**Table 5.5**: Effect of Augmentation: AUC and VR@FAR = 0.1% FAR of FRGC's fourth experiment.

| Feats. | Aug. | GMM Size | PCA Dim | Metric Learning | AUC | VR @FAR=0.1% | Z-norm AUC | Z-norm VR @FAR=0.1% |
|---|---|---|---|---|---|---|---|---|
| Gabor | ✓ | 512 | 32 | **Mah. Metric** | 98.9 | 42.9 | **99.1** | **57.4** |
| Gabor | ✓ | 512 | 32 | Joint Metric | 98.3 | 38.4 | 98.4 | 38.6 |
| Gabor | ✓ | 512 | 32 | Diag. Metric | 97.3 | 41.9 | 97.6 | 46.9 |
| Gabor | | 512 | 32 | Mah. Metric | 92.2 | 6.6 | 94.0 | 16.1 |
| Gabor | | 512 | 32 | Joint Metric | 90.8 | 7.5 | 91.9 | 9.4 |
| Gabor | | 512 | 32 | Diag. Metric | 82.2 | 6.3 | 84.5 | 8.4 |

### 5.4.4 Effect of GMM size:

In this experiment, we want to check the performance of Fisher vectors on FRGC's fourth experiment when using different size of GMM.We try using 512 and 128 sizes. GMM size parameter affects the dimensionality of Fisher vectors since their size depends on the number of vocabulary of GMM. Table 5.6 shows the evaluation results of this experiment using different size of GMM while keeping the rest setting of Fisher vectors such as PCA dimensions, Gabor response settings and augmentation fixed in the two cases. We figure out that large number of GMM leads to better results for these evaluation.

**Table 5.6**: Effect of GMM size: AUC and VR@0.1% FAR of FRGC's fourth experiment.

| Feats. | Aug. | GMM Size | PCA Dim | Metric Learning | AUC | VR @FAR=0.1% | Z-norm AUC | Z-norm VR @FAR=0.1% |
|---|---|---|---|---|---|---|---|---|
| Gabor | ✓ | 512 | 32 | **Mah. Metric** | 98.9 | 42.9 | **99.1** | **57.4** |
| Gabor | ✓ | 512 | 32 | Joint Metric | 98.3 | 38.4 | 98.4 | 38.6 |
| Gabor | ✓ | 512 | 32 | Diag. Metric | 97.3 | 41.9 | 97.6 | 46.9 |
| Gabor | ✓ | 128 | 32 | Mah. Metric | 97.9 | 27.9 | 98.3 | 41.9 |
| Gabor | ✓ | 128 | 32 | Joint Metric | 97.8 | 32.1 | 98.0 | 36.7 |
| Gabor | ✓ | 128 | 32 | Diag. Metric | 96.0 | 28.7 | 96.3 | 33.0 |

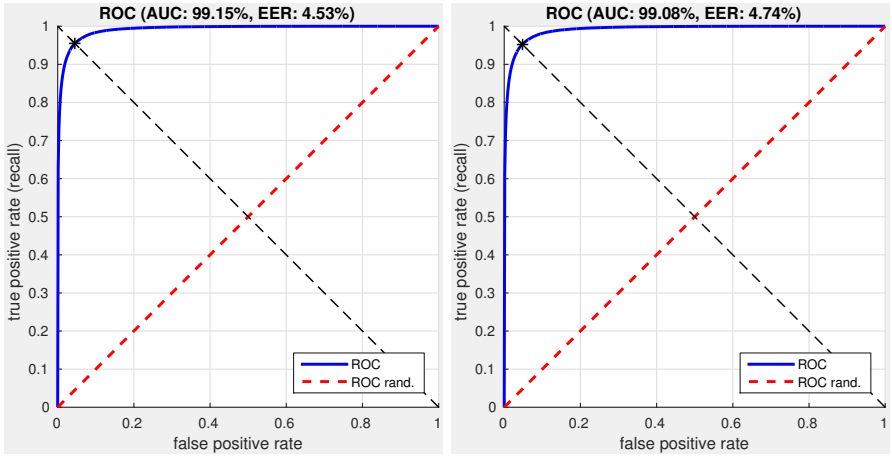
### 5.4.5 Final experiment

Following the aforementioned experiments and after exploring the best settings of Fisher vectors and Gabor responses, we select the best settings and provide their evaluation on FRGC's fourth experiment. Table 5.7 gives the results of Fisher vectors on the fourth experimenter of FRGC. For Gabor filters, 5 scales and 8 orientations are used. The obtained GMI is divided into 20 patches with $32 \times 32$ size and then down-sampled into $16 \times 32$. The resulted local descriptors dimensions is $256 \times 800$. Then PCA is applied on these descriptors to reduce the dimensionality to 32. Finally, spatial, scale and orientation information corresponding to each block is added to the local descriptors resulting in $36 \times 800$ dimensions. The GMM size is set to 512, which leads to a Fisher vector with $2PK = 2 \times 36 \times 512 = 36846$ dimensions As shown in the table, low-rank Mahalanobis metric learning has the best results in term of AUC and the FV@FAR = 0.1%. The zero-normalization of the scores improve the results, especially the VR@FAR = 0.1%, for example, when using the Mahalanobis metric learning, the VR@FAR = 0.1% = 42.87%, and after applying the zero-normalization, it increased nearly by 15% to 57.36%. This is a significant improvement over the original scores for the evaluation of FRGC's fourth experiment.

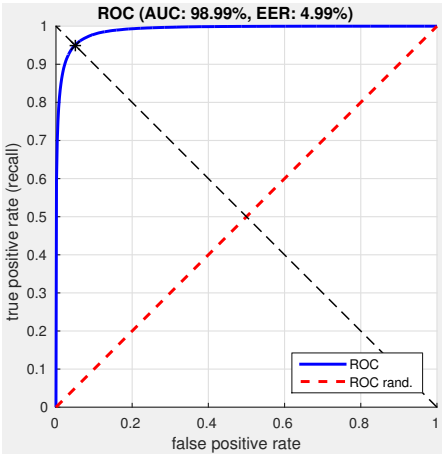
**Table 5.7**: Final Results of AUC and VR@0.1% FAR of FRGC's fourth experiment.

| Feats. | Aug. | GMM Size | PCA Dim | Metric Learning | AUC | VR @FAR=0.1% | Z-norm AUC | Z-norm VR @FAR=0.1% |
|---|---|---|---|---|---|---|---|---|
| Gabor | ✓ | 512 | 32 | **Mah. Metric** | 98.9 | 42.9 | **99.1** | **57.4** |
| Gabor | ✓ | 512 | 32 | Joint Metric | 98.3 | 38.4 | 98.4 | 38.6 |
| Gabor | ✓ | 512 | 32 | Diag. Metric | 97.3 | 41.9 | 97.6 | 46.9 |

Figures 5.5 introduces the the three ROC curves plot, where the results in table 5.7 are the average of these three ROC curves.




(a) First ROC curve plot (b) Second ROC curve plot




(c) Third ROC curve plot

**Figure 5.5**: The three ROC curves plot of the fourth experiment of FRGC using low-rank Mahalanobis metric learning.

Table 5.8 presents the results of Fisher vector on FRGC's fourth experiment for both Gabor and SIFT features. SIFT outperform Gabor filter with small difference. However, Gabor filters has less dimensions and can be computed faster than SIFT features.

**Table 5.8**: Comparison between SIFT and Gabor results of AUC and VR@0.1% FAR of FRGC's fourth experiment.

| Feats. | Aug. | GMM Size | PCA Dim | Metric Learning | AUC | VR @FAR=0.1% | Z-norm AUC | Z-norm VR @FAR=0.1% |
|---|---|---|---|---|---|---|---|---|
| Gabor | ✓ | 512 | 32 | **Mah. Metric** | 98.9 | 42.9 | **99.1** | **57.4** |
| Gabor | ✓ | 512 | 32 | Joint Metric | 98.3 | 38.4 | 98.4 | 38.6 |
| Gabor | ✓ | 512 | 32 | Diag. Metric | 97.3 | 41.9 | 97.6 | 46.9 |
| SIFT | ✓ | 512 | 64 | **Mah. Metric** | 99.26 | 51.0 | **99.4** | **63.9** |
| SIFT | ✓ | 512 | 64 | Joint Metric | 99.3 | 60.1 | 99.2 | 55.1 |
| SIFT | ✓ | 512 | 64 | Diag. Metric | 98.4 | 40.8 | 98.6 | 48.8 |

# 6. CONCLUSION AND FUTURE WORK

## 6.1 Conclusion

We introduce the *Accio* data set for analyzing the effect of aging in video face recognition methods. *Accio* features 121 characters spread over multiple age groups, and has more than 38K face tracks of which a large pool (40%) are distractor tracks. We use Fisher vector, the state-of-the-art of face track representation for the Accio dataset. The meta-data, which include the face tracks information of the dataset such as characters, track ID, age ... etc and the face tracks' representation can be downloaded at http://cvhci.anthropomatik.kit.edu/projects/mma.

In addition we carry out an study of face track retrieval and recognition across age on the dataset. Depending on whether the external and Accio data are used or not, we present three protocols: unrestricted, restricted, and free-for-all. We present two primary tasks as a benchmark for the face track retrieval evaluation: within and across movie face track retrieval. During our study, the baseline experiments of both tasks (within and across movies evaluation) for face track retrieval and recognition we see a steady decline in the retrieval performance as the age gap between the query and database movies increases.

In order to improve the baseline results we benefit from three metric learning techniques: (i) *low-rank Mahalanobis metric learning*, (ii) *joint metric learning*, and (iii) *diagonal metric learning*. These techniques serve two majors: (i) dimensionality reduction and (ii) more discriminative representation in lower space. In addition, during the evaluation on Accio dataset, we have two different usage of metric learning based on the pooled training data: (i) *metric learning on pairwise movies*, and (ii) *metric learning on all movies training data*. In both scenarios, we observe an improvement over the baseline results in case of using joint and low-rank Mahalanobis metric learning in term of both mAP and precision @k. Their results show clearly the ability of these metric learning approaches to learn the feature similarity of the facial

changes due to the aging factor. However, diagonal metric learning, produces lower results than baseline since it is basic metric learning that can not capture the variation of face track representation due to the facial changes across age.

We carry out a supervised learning by linear SVMs in order to assess the efficiency of face track recognition across age on this dataset. Similarly we have two evaluation tasks: within and across movies face recognition. As the retrieval task, we observe a decline in the system performance as the gap between the training and testing movies increases and the best results are obtained in case of applying training and test on the same movies.

As a part of our work on face track retrieval and recognition across age, we evaluate the efficiency of Fisher vectors using different parameters. We implement multiple experiments, using FRGC dataset and Gabor filters. We use variant parameters of Fisher vectors and face image in our experiments such as different face alignment, various number of GMM and PCA sizes, and the augmentation of spatial, scale and orientation information to the local descriptors. We conclude our study by seeing that, Fisher vector method works well without any regard to face alignment since it can capture the important features of the face with out needing to specific alignment. The augmentation of spatial, scale and orientation to the Fisher vectors clearly affect the results of Fisher vectors since it helps the technique to learn the spatial and visual distribution of the local features. Finally, we notice that, Fisher vectors works better with low dimensions of PCA since it makes the representation more discriminative.

## 6.2 Future Work

In the future work, our purpose is to obtain an age invariant face track retrieval and recognition on Accio dataset using Transfer Learning. One of the proposed technique is instance transfer which includes copying training set tracks directly or smart copy such that the accuracy of across movies evaluation improves. In addition, model transfer can be applied where we get scores from all models and build a top model on this to get better performance. Domain adaptation of the features is also suggested which combine all movies into one domain to learn the feature similarity of face tracks regardless of the facial changes due to age progression.

In addition, we plan to study the face track retrieval and recognition across age based on age groups. For example, the experiments would be implemented explicitly on certain age groups such as young ages which are dominant in Accio dataset. Furthermore, in Accio dataset age invariant learning and representation are promising since the aging factor on the dataset has been proved. The work will be accompanied with more research and work on age invariant face representation for certain facial landmarks and ages, and learning that captures the facial feature changes due to age progression across movies as suggested in the literature.

## REFERENCES


[1] **Kowdle, A. and Chen, T.** (2012). Learning to Segment a Video to Clips Based on Scene and Camera Motion, *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ECCV'12, Springer-Verlag, Berlin, Heidelberg, pp.272–286, http://dx.doi.org/10.1007/978-3-642-33712-3_20. xix, 13

[2] **Fröba, B. and Ernst, A.** (2004). Face Detection with the Modified Census Transform, *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, FGR' 04, IEEE Computer Society, Washington, DC, USA, pp.91–96, http://dl.acm.org/citation.cfm?id=1949767.1949786. xix, 14, 15

[3] **Simonyan, K., Parkhi, O.M., Vedaldi, A. and Zisserman, A.** (2013). Fisher Vector Faces in the Wild, *British Machine Vision Conference*. xix, 3, 5, 9, 10, 17, 18, 22, 31, 60

[4] **Phillips, P., Flynn, P., Scruggs, T., Bowyer, K. and Worek, W.** (2006). Preliminary Face Recognition Grand Challenge Results, *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pp.15–24. xx, 10, 59, 60

[5] **Gao, H., Ekenel, H., Fischer, M. and Stiefelhagen, R.** (2010). Multi-resolution Local Appearance-Based Face Verification, *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp.1501–1504. xx, 4, 9, 10, 19, 62

[6] **Bäuml, M., Tapaswi, M. and Stiefelhagen, R.** (2013). Semi-supervised Learning with Constraints for Person Identification in Multimedia Data, *CVPR*. 1, 10, 11, 14, 16, 30

[7] **Everingham, M., Sivic, J. and Zisserman, A.** (2006). "Hello ! My name is ... Buffy" – Automatic Naming of Characters in TV Video, *BMVC*. 1, 10, 11, 34

[8] **Parkhi, O.M., Simonyan, K., Vedaldi, A. and Zisserman, A.** (2014). A Compact and Discriminative Face Track Descriptor, *CVPR*. 1, 11, 17, 31

[9] **Feris, R., Bobbitt, R., Brown, L. and Pankati, S.** (2014). Attribute-based People Search: Lessons Learnt from a Practical Surveillance System, *ICMR*. 1

[10] FG-NET Aging Database. 2, 10, 33

[11] **Ricanek, K. and Tesafaye, T.** (2006). MORPH: A longitudinal image database of normal adult age-progression, *IEEE FG*. 2, 10, 33

[12] **Chen, B.C., Chen, C.S. and Hsu, W.H.** (2014). Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval, *ECCV*. 2, 10, 11, 33

[13] **Huang, G.B., Ramesh, M., Berg, T. and Learned-Miller, E.** (2007). Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, **Technical Report07-49**, Univ. of Massachusetts, Amherst. 3, 10

[14] **Wolf, L., Hassner, T. and Maoz, I.** (2011). Face Recognition in Unconstrained Videos with Matched Background Similarity, *CVPR*. 3, 10

[15] **Chen, D., Cao, X., Wang, L., Wen, F. and Sun, J.** (2012). Bayesian Face Revisited: A Joint Formulation, *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ECCV'12, Springer-Verlag, Berlin, Heidelberg, pp.566–579, http://dx.doi.org/10.1007/978-3-642-33712-3_41. 3, 9, 23

[16] **Chen, D., Cao, X., Wen, F. and Sun, J.** (2013). Blessing of Dimensionality: High-Dimensional Feature and Its Efficient Compression for Face Verification, *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp.3025–3032. 3, 23

[17] **Weinberger, K.Q., Blitzer, J. and Saul, L.K.** (2006). Distance metric learning for large margin nearest neighbor classification, *In NIPS*, MIT Press. 3, 10

[18] **Viola, P. and Jones, M.J.** (2004). Robust Real-Time Face Detection, *Int. J. Comput. Vision*, ***57***(2), 137–154, http://dx.doi.org/10.1023/B:VISI.0000013087.49260.fb. 9, 14

[19] **Bradski, G.** *Dr. Dobb's Journal of Software Tools*. 9

[20] **Küblbeck, C. and Ernst, A.** (2006). Face Detection and Tracking in Video Sequences Using the Modifiedcensus Transformation, *Image Vision Comput.*, ***24***(6), 564–572, http://dx.doi.org/10.1016/j.imavis.2005.08.005. 9

[21] **Xiong, X. and De la Torre Frade, F.** (2013). Supervised Descent Method and its Applications to Face Alignment, *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. 9

[22] **Everingham, M., Sivic, J. and Zisserman, A.** (2009). Taking the Bite out of Automatic Naming of Characters in TV Video, *Image and Vision Computing*, ***27***(5). 9

[23] **Ahonen, T., Hadid, A. and Pietikainen, M.** (2006). Face Description with Local Binary Patterns: Application to Face Recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, ***28***(12), 2037–2041, http://dx.doi.org/10.1109/TPAMI.2006.244. 9

[24] **Ekenel, H.K. and Stiefelhagen, R.** (2005). Local Appearance based Face Recognition Using Discrete Cosine Transform, *13th European Signal Processing Conference (EUSIPCO 2005*. 9

[25] **Taigman, Y., Wolf, L. and Hassner, T.** (2009). Multiple One-Shots for Utilizing Class Label Information, *The British Machine Vision Conference (BMVC)*, http://www.openu.ac.il/home/hassner/projects/multishot. 9

[26] **Wolf, L., Hassner, T. and Taigman, Y.** Y.: Descriptor based methods in the wild, *In: Faces in Real-Life Images Workshop in ECCV. (2008) (b) Similarity Scores based on Background Samples*. 9

[27] **Wolf, L., Hassner, T. and Taigman, Y.** (2010). Similarity Scores Based on Background Samples, *Proceedings of the 9th Asian Conference on Computer Vision - Volume Part II*, ACCV'09, Springer-Verlag, Berlin, Heidelberg, pp.88–97, http://dx.doi.org/10.1007/978-3-642-12304-7_9. 9

[28] **Guillaumin, M., Verbeek, J. and Schmid, C.** (2009). Is that you? Metric learning approaches for face identification, *Computer Vision, 2009 IEEE 12th International Conference on*, pp.498–505. 9

[29] **Setty, S. and et al.** (2013). Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations, *NCVPRIPG*. 10

[30] **Sivic, J., Everingham, M. and Zisserman, A.** (2005). Person spotting: video shot retrieval for face sets, *CIVR*. 10, 11

[31] **Everingham, M. and Zisserman, A.** (2005). Identifying individuals in video by combining 'generative' and discriminative head models, *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pp.1103–1110 Vol. 2. 10

[32] **Tapaswi, M., Bäuml, M. and Stiefelhagen, R.** (2012). "Knock! Knock! Who is it?" probabilistic person identification in TV-series., *CVPR*, IEEE Computer Society, pp.2658–2665. 10

[33] **Sivic, J., Everingham, M. and Zisserman, A.** (2009). Who are you? - Learning person specific classifiers from video, *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp.1145–1152. 10

[34] **Wolf, L., Hassner, T. and Maoz, I.** (2011). Face recognition in unconstrained videos with matched background similarity, *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp.529–534. 10

[35] **Krueger, V., Gross, R. and Baker, S.** (2002). Appearance-based 3-D Face Recognition from Video, *Proceedings of the German Symposium on Pattern Recognition (DAGM)*. 11

[36] **Zhou, S., Krueger, V. and Chellappa, R.** (2002). Face recognition from video: a CONDENSATION approach, *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pp.221–226. 11

[37] **Li, H.**, **Hua, G.**, **Lin, Z.**, **Brandt, J. and Yang, J.** (2013). Probabilistic Elastic Matching for Pose Variant Face Verification, *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp.3499–3506. 11

[38] **Fischer, M.**, **Ekenel, H.K. and Stiefelhagen, R.** (2010). Person re-identification in TV series using robust face recognition and user feedback, *Multimedia Tools and Applications*. 11, 22

[39] **Li, Y.**, **Wang, R.**, **Cui, Z.**, **Shan, S. and Chen, X.** (2014). Compact Video Code and Its Application to Robust Face Retrieval in TV-Series, *BMVC*. 11

[40] **Geng, X.**, **Zhou, Z.H. and Smith-Miles, K.** (2007). Automatic Age Estimation Based on Facial Aging Patterns, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, ***29***(12), 2234–2240. 11

[41] **Guo, G.**, **Mu, G.**, **Fu, Y. and Huang, T.** (2009). Human age estimation using bio-inspired features, *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp.112–119. 11

[42] **Kwon, Y.H. and Lobo, N.D.V.** (1999). Age Classification from Facial Images, *In Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp.762–767. 11

[43] **Lanitis, A.**, **Draganova, C. and Christodoulou, C.** (2004). Comparing Different Classifiers for Automatic Age Estimation, *Trans. Sys. Man Cyber. Part B*, ***34***(1), 621–628, http://dx.doi.org/10.1109/TSMCB.2003.817091. 11

[44] **Montillo, A. and Ling, H.** (2009). Age regression from faces using random forests, *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pp.2465–2468. 11

[45] **Yan, S.**, **Wang, H.**, **Tang, X. and Huang, T.** (2007). Learning Auto-Structured Regressor from Uncertain Nonnegative Labels, *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp.1–8. 11

[46] **Zhou, S.**, **Georgescu, B.**, **Zhou, X.S. and Comaniciu, D.** (2005). Image based regression using boosting method, *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pp.541–548 Vol. 1. 11

[47] **Wang, J.**, **Shang, Y.**, **Su, G. and Lin, X.** (2006). Age simulation for face recognition, *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pp.913–916. 11

[48] **Du, J.X.**, **Zhai, C.M. and Ye, Y.Q.** (2011). Face Aging Simulation Based on NMF Algorithm with Sparseness Constraints., *ICIC*, volume6839, Springer, pp.516–522. 11

[49] **Lanitis, A.**, **Taylor, C. and Cootes, T.** (2002). Toward automatic simulation of aging effects on face images, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, ***24***(4), 442–455. 11

[50] **Park, U., Tong, Y. and Jain, A.** (2010). Age-Invariant Face Recognition, *T-PAMI*, ***32***(5), 947–954. 11

[51] **Suo, J., Chen, X., Shan, S. and Gao, W.** (2009). Learning long term face aging patterns from partially dense aging databases, *Computer Vision, 2009 IEEE 12th International Conference on*, pp.622–629. 11

[52] **Suo, J., Zhu, S.C., Shan, S. and Chen, X.** (2010). A Compositional and Dynamic Model for Face Aging, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, ***32***(3), 385–401. 11

[53] **Tsumura, N., Nakaguchi, T. and Ojima, N.**, Image-Based Control of Skin Melanin Texture. 11

[54] **Geng, X., Zhou, Z.H. and Smith-Miles, K.** (2007). Automatic Age Estimation Based on Facial Aging Patterns, *T-PAMI*, ***29***(12), 2234–2240. 11

[55] **Lanitis, A., Draganova, C. and Christodoulou, C.** (2004). Comparing different classifiers for automatic age estimation, *IEEE Trans Syst Man Cybern: B Cybern*, ***34***(1), 621–628. 11

[56] **Ramanathan, N. and Chellappa, R.** (2006). Modeling Age Progression in Young Faces, *CVPR*. 11

[57] **Klare, B. and Jain, A.K.** (2011). Face recognition across time lapse: On learning feature subspaces, *Intl. Joint Conf. on Biometrics*. 11

[58] **Li, Z., Park, U. and Jain, A.K.** (2011). A Discriminative Model for Age Invariant Face Recognition, *IEEE Trans on Inf. Forensics Security*, ***6***(3). 11

[59] **Ling, H., Soatto, S., Ramanathan, N. and Jacobs, D.** (2010). Face Verification Across Age Progression Using Discriminative Methods, *IEEE Trans on Inf. Forensics Security*, ***5***(1), 82–91. 11

[60] **Gong, D., Li, Z., Lin, D., Liu, J. and Tang, X.** (2013). Hidden Factor Analysis for Age Invariant Face Recognition, *ICCV*. 12

[61] **Lienhart, R.** (2001). Reliable Transition Detection In Videos: A Survey and Practitioner's Guide, *International Journal of Image and Graphics (IJIG) Vol. 1, No. 3, pp. 469-486*. 13

[62] **Yusoff, Y., Christmas, W.J. and Kittler, J.** (1998). A Study on Automatic Shot Change Detection, *Multimedia Applications, Services and Techniques - ECMAST '98, Third European Conference, Berlin, Germany, May 26-28, 1998, Proceedings*, pp.177–189, http://dx.doi.org/10.1007/3-540-64594-2_94. 13

[63] **Nummiaro, K., Koller-meier, E. and Gool, L.V.** (2002). A Color-based Particle Filter, pp.53–60. 16

[64] **Cao, X., Wei, Y., Wen, F. and Sun, J.** (2013). Face Alignment by Explicit Shape Regression, *International Journal of Computer Vision*, http://research.microsoft.com/apps/pubs/default.aspx?id=206636. 16

**CURRICULUM VITAE**

**Name Surname: Esam Ghaleb**

**Place and Date of Birth: Taiz, Yemen - 04/01/1989**

**E-Mail: ghalebe@itu.edu.tr**

**B.Sc.: Istanbul Technical University**

**M.Sc.: Istanbul Technical University**

**List of Publications**

Ugur Demir, **Esam Ghaleb**, and Hazım Kemal Ekenel, "A Face Recognition Based Multiplayer Mobile Game Application", *Artificial Intelligence Applications and Innovations (AIAI 2014)*

Abdullah Aydeger, **Esam Ghaleb**, Sema Oktuğ, "AN ENERGY EFFICIENT ROUTING TECHNIQUE AND IMPLEMENTATION IN WSNs", *IEEE 22. Sinyal İşleme ve İletişim Uygulamaları Kurultayı(SIU 2014)*

**PUBLICATIONS/PRESENTATIONS ON THE THESIS**

▪ **Esam Ghaleb**, Makarand Tapaswi, Ziad Al-Halah, Hazım Kemal Ekenel, Rainer Stiefelhagen, "Accio: A Data Set for Face Track Retrieval in Movies Across Age", *ACM International Conference on Multimedia Retrieval (ICMR 2015)*

**SCHOLARSHIPS**

**[1]** 2007 – 2013: Scholarship of "Ministry of High Education in Yemen" and "Ministry of High Education in Turkey" in scope of supporting succesful Yemeni students

**[2]** 2013 - ... ITU Scholarship as project assistance under the scope of "Collaborative Annotation of multi-MOdal, multI-Lingual and multi-mEdia documents (CAMOMILE)" European Project