UNIVERSITY OF TARTU

FACULTY OF SCIENCE AND TECHNOLOGY

Institute of Technology

Computer Engineering

Tõnis Uiboupin

Super Resolution and Face Recognition Based People Activity Monitoring Enhancement Using Surveillance Camera

Master's Thesis (30 ECTS)

Supervisors: Assoc. Prof. Gholamreza Anbarjafari

Pejman Rasti, M.Sc.

Tartu 2016

Abstract/Resümee

Super Resolution and Face Recognition Based People Activity Monitoring Enhancement Using Surveillance Camera

Due to importance of security in the society, monitoring activities and recognizing specific people through surveillance video camera is playing an important role. One of the main issues in such activity rises from the fact that cameras do not meet the resolution requirement for many face recognition algorithms. In order to solve this issue, in this work we are proposing a new system which super resolve the image. First, we are using sparse representation with the specific dictionary involving many natural and facial images to super resolve images. As a second method, we are using deep learning convulutional network. Image super resolution is followed by Hidden Markov Model and Singular Value Decomposition based face recognition. The proposed system has been tested on many well-known face databases such as FERET, HeadPose, and Essex University databases as well as our recently introduced iCV Face Recognition database (iCV-F). The experimental results shows that the recognition rate is increasing considerably after applying the super resolution by using facial and natural image dictionary. In addition, we are also proposing a system for analysing people movement on surveillance video. People including faces are detected by using Histogram of Oriented Gradient features and Viola-jones algorithm. Multi-target tracking system with discrete-continuouos energy minimization tracking system is then used to track people. The tracking data is then in turn used to get information about visited and passed locations and face recognition results for tracked people.

CERCS - T111 Imaging, image processing

Keywords: Super Resolution, Deep Learning, Surveillance Videos, Face Recognition, Hidden Markov Model, Singular Value Decomposition, Human Tracking, Histogram of Oriented Gradients.

Valvekaameratel põhineva inimseire täiustamine pildi resolutsiooni parandamise ning näotuvastuse abil

Tänapäeva ühiskonnas omab turvalisus olulist rolli. Sellest tulenevalt on olulisel kohal ka valvekaameratel põhinev inimeste jälgimine ning soovitud isikute tuvastamine. Üheks olulisemaks probleemiks selle juures on videopildilt eraldatud nägude madal resolutsioon. Enamik näotuvastualgoritme ei suuda madala resolutsiooniga nägude põhjal piisava täpsusega inimesi ära tunda. Selle probleemi lahendamiseks on antud töös välja pakutud uus süsteem, mis suurendab näopiltide resolutsiooni kasutades super-resolutsiooni tehnikaid. Esiteks on kasutatud hajusesitusel põhinevat superresolutsiooni meetodit (*Sparse Representation based Super Resolution*), mis kasutab madala resolutsiooniga pildile vastava kõrgema resolutsiooniga pildi konstrueerimiseks madala ja kõrge resolutsiooniga pildifragmentide kogu ehk sõnastikku. Teiseks, konvolutsioonaalsete võrkude süvaõppel põhinevat superresolutsiooni meetodit (*Convolutional Neural Network based Super Resolution*). Pärast sisendpildi resolutsiooni parandamist teostatakse näotuvastus kasutades varjatud Markovi mudeleid (*Hidden Markov Models*) koos singulaarsete väärtuste dekompositsiooniga (*Singular Value Decomposition*). Süsteemi on katsetatud mitmetel tuntud nägude andmebaasidel nagu FERET, HP ja Essexi ülikooli andmebaas, lisaks ka töö raames loodud uuel nägude andmebaasil iCV-F. Eksperimentide tulemused näitasid, et näotuvastuse tulemus paranes tunduvalt pärast pildi resolutsiooni suurendamist. Lisaks on välja pakutud uus süsteem analüüsimaks videovoolt inimeste liikumist. Inimeste ja nende nägude leidmiseks on kasutatud orienteeritud kallete histogrammi (*Histogram of Oriented Gradient*) ning Viola-Jones algoritmi. Tuvastatud inimeste asukohtade andmed kasutatakse ära inimeste jälgimise süsteemis. Selleks on kasutatud mitme sihtmärgi jälgimise süsteemi, mis kasutab diskreet-pidevat energia vähendust (*Multi-target tracking system with discrete-continuouos energy minimization*). Töö käigus täiendati olemasolevat jälgimissüsteemi. Täienduse tulemusena väljastab see informatsiooni jälgitud sihtmärkide teekonna kohta läbitud ning külastatud punktide järjestuse kaudu, lisaks ka näotuvastuse tulemused.

CERCS - T111 Pilditehnika

Märksõnad: Superresolutsioon, Süvaõpe, Valvekaamera Video, Näotuvastus, Varjatud Markovi Mudel, Singulaarsete Väärtuste Dekompositsioon, Inimseire, Orienteeritud Kallete Histogramm.

Contents

List of Figures

List of Tables

Abbreviations, constants, generic terms

a - Data association of detection

B - Vector with n dimensions

c - Number of channels of an image

CART - Classification and Regression Tree Analysis

CNN - Convolutional Neural Network

CR - Compensation ratio

d - Person in database

D_l, D_h - Dictionaries of low and high resolution image patches

DARPA - Defense Advanced Research Products Agency

DCT - Discrete Cosine transform

E - Energy

F - Mapping

f - Filter

FERET - Facial recognition technology database

FM - Foreground mask

G - Set of positive FR results for a trajectory

H - Height

HMM - Hidden Markov Models

HOG - Histogram of Oriented Gradient

HP - HeadPose database

HR - high resolution

I - Image

i - Counter of elements in an array

ICA - Independent Component Analysis

iCV-F - iCV Face Recognition database

L - Length

l - Threshold of converting image to binary image

LAM - Linear Associative Memories

LD - Label of detection

LDA - Linear discriminant analysis

LHR - low and high resolution

LR - low resolution

m - Person on a test image

MAP - Maximum a Posterior

MCA - Morphological Component Analysis

MLP - Multi Layer Perceptron

N**,** M - Dimensions of high- and low-resolution patches

n - Number of different values in an array

NN - Neural Networks

p - Image patch

PCA - Principal components analysis

r - Detection

ReLU - Rectified Linear Unit

RR - Recognition rates

S - Downsampling operator

SF - Spatial size of filter

SR - Super Resolution

SVD - Singular Value Decomposition

SVM - Support vector machines

T - Set of target trajectories

TH - Target hypotheses

U, Σ **and V** - Matrices used by SVD

W - Width

X, x_i, Q_i - Observation vector for HMM, quantized value of vector, number of distinct levels quantized

XML - Extensible Markup Language

Y - Set of filters

Z - Normalization factor

α_l,α_h - Representations of LHR patch-pair in D_l and D_h

λ - HMM face model

$\hat{\lambda}$ - Sparsity regularization parameter

Υ - operations

$.\times$ - Element wise operation

1 Introduction

1.1 Problem overview

Today, world security standards in crowded areas like airports, metro, shopping malls have been highly increased. One way for avoiding threats is to use video surveillance. Systems like CCTV have been around for quite a while, but has required a lot of manual work to check the live feed or saved data.
One way to make this more efficient is to deploy automated system to detect and recognise people from the video stream. This requires finding the face and applying face recognition on it. Such automation is complicated by multiple factors. One of the main problems is low image resolution, because of cameras often covering large areas to reduce cost. A way to reduce such an effect is to apply image super resolution (SR).

Another type of information available through surveillance video is the path of movement of the people. Getting this information manually requires unreasonable amount of work. In order to automate this a system for detecting people from video stream is required. Additionally these detections need to be linked together by tracking system to get the trajectories of moving people.

1.2 Goals

In the first part of this thesis we are testing the effect of enhancing face image resolution by applying SR to race recognition performance. State-of-the-art sparse representation [1] and Convolutional Neural Network (CNN) based SR algorithms [2] are investigated in improving recognition accuracies of the state-of-the-art face recognition algorithm of [3]. To the extent of our knowledge, face recognition in different resolutions has not been studied extensively, except in [4] in which optical flow based SR is used to enhance accuracies of face recognition. However, our proposed system is the first one in which SR algorithms have been employed for improving quality of low resolution (LR) input facial images before face recognition. We show in this thesis that such SR algorithms produce high resolution (HR) details that are not necessarily recovered by simple upscaling algorithms, like bicubic interpolation. It is shown in this thesis that results of the sparse representation and deep learning-based SR produce images that are of better quality compared to the input LR images. We show that employing such higher resolution images improves the recognition accuracy of a state-of-the-art face recognition algorithm.

In the second part of this thesis a system is implemented for testing a scenario of tracking people. Histogram of Oriented Gradient (HOG) features and Viola-jones algorithm [5] are used for detecting people and faces. Multi-target tracking system with discrete-continuouos energy minimization tracking system [6] is then used to track people. The resuting data of the tracking system is then used to get information about visited and passed locations. Face recognition results from detections together with tracking data is finally used to recognize tracked people.

The remainder of this thesis is organized as follows. Chapter 2 contains literature review on image super-resolution. An overview of state-of-the-art face recognition systems and image pre-processing is given in Chapter 3. A detailed overview of the proposed system is presented in Chapter 4. In Chapter 5, the outcome of experimental results are presented. Finally, Chapter 6 concludes the work.

2 Super resolution

Image super resolution (SR) techniques aim at enhancing the resolution of images acquired by low resolution (LR) sensors, while also minimizing added visual artifacts. Such techniques are expected to enable overcoming limitations of LR imaging such as surveillance cameras. This field of study has been very attractive research topic for more than two decades now. SR has many practical applications in helping to solve real world problems in different fields, including satellite and aerial imaging, medical image processing, text image analysis, sign and number plates reading, biometrics recognition, facial image analysis and there are many more. Because of this, many research papers have been written, each proposing new SR method for specific purpose. There are different ways of classifying SR algorithms. Some use spatial domain, some frequency domain. Some use just one image, some use multiple images. Plus there is number of different image reconstruction methods used. In this work we have focused on single image SR in spatial domain. SR algorithms using frequency domain or multy-image approach have not been proven to be very effective in this area. The field of single image SR has been shown considerable attention [1, 7–9]. Figure **2.1** shows the proposed taxonomy for SR algorithms by the authors of [10] and also some different single image based algorithms.

Interpolation has been widely adopted in many SR techniques [11–13]. In past five to ten years a wide range of new approaches have been suggested which improve conventional linear interpolators.

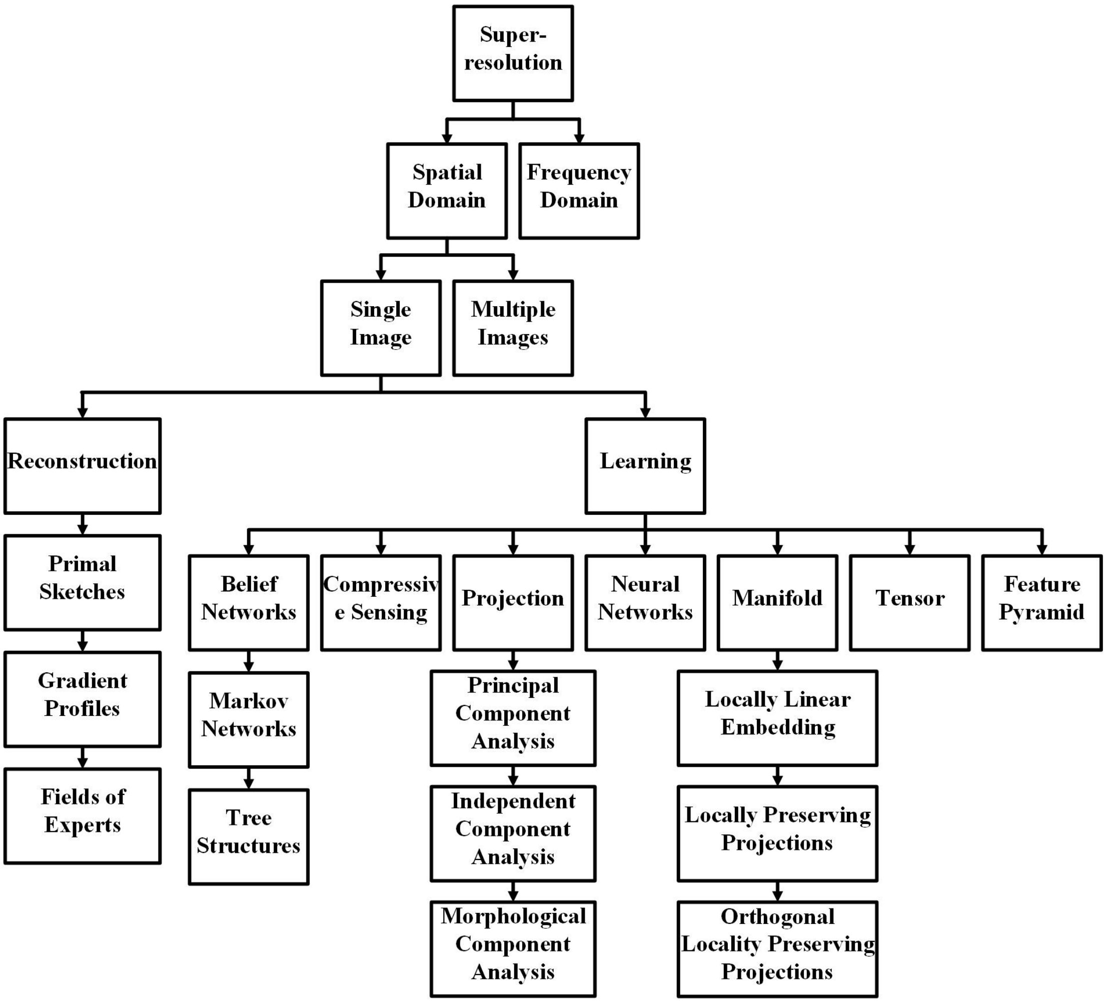

Figure 2.1: Single image SR algorithms.

2.1 Learning based Single Image SR

Learning based Single Image SR or Hallucination algorithms were first introduced in 1985 [14]. These algorithms first learn the relationship between some HR examples and their LR counterparts in training step. Then this knowledge is used to reconstruct HR image from LR images. Examples for training are picked from specific class such as fingerprints, face images, etc.

In training step of Feature Pyramids algorithm, a Gaussian resolution pyramid is produced by first down-sampling each HR face image and then blurring several times. Laplacian and Feature pyramids are then generated from these Gaussian pyramids. Once the system has been trained, the most similar LR image to LR input image/patch is found from all the pyramids. Then the relationships between the found LR image and its corresponding HR image are used to predict the HR details of the LR input image [15–18].

Belief Network algorithms use a belief network such as a Markov Network [19] or a tree structure [20] for learning the relationship between the LR and corresponding HR images in the training step.

For learning the a priori term of the employed Maximum a Posteriori (MAP) algorithm, some SR algorithms use methods that are based on projection. Such SR algorithms include Principal Component Analysis (PCA) [21], Independent Component Analysis (ICA) [22] and Morphological Component Analysis (MCA) [23].

Neural Networks (NN) based SR algorithms share the same concept as belief-network based methods. Some of such networks are Linear Associative Memories (LAM) with single [24] and dual associative learning [25], Hopfield NN [26], Probabilistic NN [27], Integrated Recurrent NN [28] and Multi Layer Perceptron (MLP) [29]. Additionally, Dong [2] proposed a deep convolutional neural network (CNN) based single image super resolution method and showed that the traditional sparse-coding-based algorithm can also be seen as a kind of deep convolutional network. The end-to-end mapping between LR images and HR images was optimized in Dong's SR method, which achieves excellent reconstruction performance.

In Manifold based methods, HR and LR images are assumed to form manifolds which have similar local geometries in two distinct feature spaces [30]. These methods, like PCA, are usually used for dimensionality reduction. Manifold based methods include generally following steps: training step where HR and LR manifolds are generated using HR images and their corresponding LR counterparts and testing step where input image is divided into a set of LR patches to reconstruct HR patches. For each LR patch, in testing step *k*-nearest neighbor patches are found from LR manifold. These *k*-nearest neighbors are then used to calculate weights for reconstructing the LR patch. Involved weights and neighbors are then used to reconstruct the HR patch by finding the corresponding objects in the HR manifold.

Tensors are multilinear equivalents of vectors (first order) and matrices (second order). In Tensor analysis the mappings between multiple factor spaces for linear methods, such as PCA, are studied. It can be considered as generalized extentsion of traditional linear methods [31].

Compressive Sensing SR algorithms include methods that use sparse coding. Techniques in [1] and [7] are based on sparsely representing low and high resolution (LHR) patch-pairs in a dictionary pair, namely dictionary low (D_l) and dictionary high (D_h). The main idea behind this approach is that each LHR patch-pair is sparsely represented over the dictionaries where the resulting representations α_l,α_h have pre-specified correspondence. Sparse representation invariance [1, 7] is a common assumption here. The idea is that patches in an LHR pair have the same sparse representation over the LHR dictionary pair α_l=α_h. Joint learning of the dictionary pair is used to get meaningful recovery of the HR patch. The authors of [1, 7] have suggested that for recovering the HR patches, first a dictionary D_l that best fits the LR patches should be learned, and then a dictionary D_h that works best with the resulting coefficients α_l. In [9], an SR algorithm based on sparse representation of patch-pairs over a dictionary pair was introduced. However no invariance assumption is made. Instead a parametric model which captures the statistical dependencies between the sparsity patterns of the LHR coefficients and between the corresponding nonzero coefficients is suggested. Prediction of α_l from α_h is done using the MMSE estimator, which arises directly from the model and has a closed form formula. This model does not require the dictionary pair to be strictly aligned or even to be of the same size. This removes any restrictions of the LR dictionary. Because of this, very small orthogonal dictionaries are used for the LR patches, which help to reduce the computational cost of the scale-up scheme.

The proposed method in [8] uses self-similarity of image patches within and across different scales in the same image in single image SR.

3 Face Recognition

3.1 Pre-processing

Viola-Jones algorithm [5] was used in order to find faces from images and video frames. The authors of this algorithm, introduced a novel real-time face detection technique and it is one of most commonly used face detector techniques today. The algorithm uses feature-based approach where classifier is trained for features selected by AdaBoost algorithm. Face detection is done by scanning test image with rectangular window at different scales and positions. The regions that pass the classifier are declared as faces. Viola and Jones also introduced a new image representation called an Integral Image. This can be computed from an image using few operations per pixel, but once computed, it allows for very fast feature evaluation. Furthermore, this technique uses learning of several classifiers which have been cascaded, instead of single classifier which would require computing all features for all scanning windows in the image. Cascading is first done for most simple classifiers and proceeded to more complex ones. In order to detect a region of an image as face, it needs to pass through all classifiers. Rejection at any stage means the region is discarded and will not be processed in later stages. Rejection of non-face regions as early as possible means it won't reach more complex classifiers and therefore speeds up the whole detection process. Figure **3.1** shows red bounding box around face detected from image by Viola-Jones algorithm. Usually some illumination enhancement is also done in pre-processing for face recognition, but as we have videos with good enough illumination, we did not do it here.

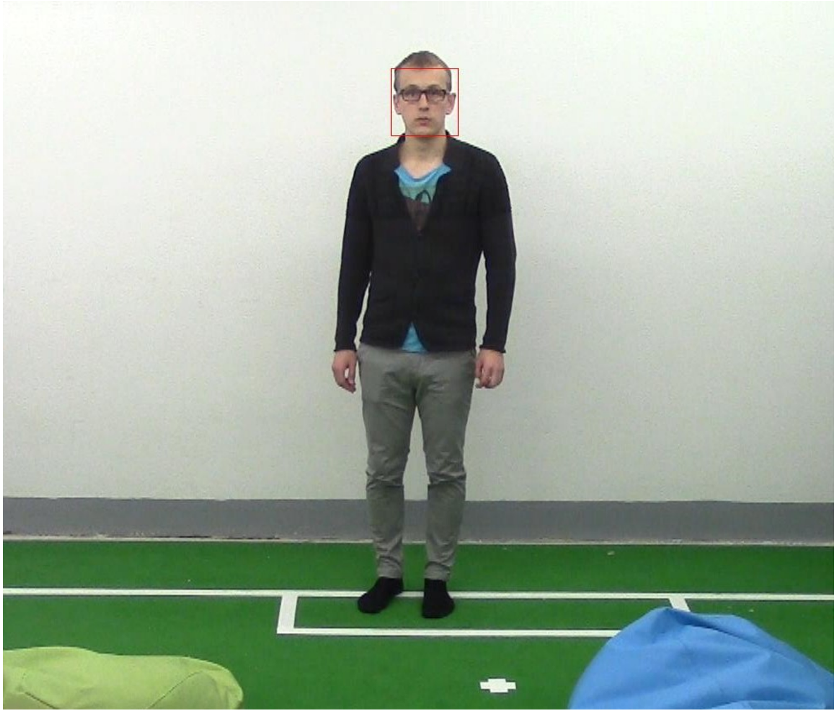

Figure 3.1: Face detected by Viola-Jones algorithm.

3.2 State-of-the-art in face recognition systems

As with image super resolution, there are also many different approaches proposed and developed for face recognition. Most relevant ones to this work are more explained here.

Principal components analysis (PCA) is an appearance-based technique which results in eigenfaces [32, 33]. The idea behind PCA mechanism is dimensionality reduction. It is done by reflecting eigenvectors of the data covariance matrix. Its objective is to gather a set of basis functions that are mutually orthogonal. These functions accumulate the directions of concentrated variance in the data. The coefficients of basis functions are pairwise decorrelated. A drawback of this method is that it is sensitive to lightning conditions and a number of images in different positions is required to get desired output.

Linear discriminant analysis (LDA) [34, 35] is another appearance-based technique.

Here discriminatory information is encoded in a linear separable space. Bases of which don't have to be necessarily orthogonal. Fishers linear discriminant criterion [34] is applied. As a result images of the same class are grouped together and images of different classes are separated from each other. For classifying the training images are projected into a subspace similarly to eigenspace projection. The test images are then projected into the same subspace. Identification of test images is done using a similarity measure. Projected test image is compared to each projected training image, and the test image is identified by the identity of the closest training image. One way of calculating minimum distance is using the Euclidian distance method.

Support vector machines (SVM) is intended as a technique for pattern recognition. It has also been used in fece recognition systems. The Authors of [36] proposed to use SVMs for learning the discrimination functions between each face pair once the features are extracted. Face images are here represented by using eigenfaces. A bottom-up binary tree structure is used for multi-class classification. Each class are in the bottom level of the binary tree in the beginning. Then, all classes are put into pairs and "winner" of the comparison of each pair moves to next level.This process is then repeated until the unique class appear on the top of the tree.

Hidden Markov Models (HMM) based approaches have been widely used in recent years [3]. Samaria and Harter [37] suggested to model facial features from frontal images by a top-bottom HMM. Kohir and Desai [38] proposed using DCT coefficients as features in a top-down HMM, which increased face recognition performance. A pseudo 2D representation of the face by using a combination of left-right models arranged in an ordered sequence [39] were reported to achieve higher recognition rates. More recently HMM based face recognition methods, like Pseudo 2D HMM with DCT coefficients features [40], Singular Value Decomposition (SVD) coefficients [41], and with Wavelet coefficients features [42], have been proposed [43].

Singular Value Decomposition (SVD) can be used to represent algebraic properties of an image [44]. Singular values of a data matrix give information about the noise level, the entry, the rank of the matrix, etc. and reflect some features of the patterns of the matrix.

4 Proposed system

4.1 HMM based face recognition

In this work, first, the Viola-Jones [5] algorithm was used to extract only faces from each image to reduce the effects of background and clothing to face recognition results. After acquiring HR, LR and SR images, face recognition algorithm was applied on images from each database. We used Hidden Markov Model (HMM)-based face recognition system [3]. It uses one dimensional Discrete HMM as classifier and Singular Values Decomposition (SVD) coefficients as features for face recognition. A seven-state HMM is used for modelling face configuration which takes into account following face regions: hair, forehead, eyebrows, eyes, nose, mouth and chin.

Each database used in our work has 10 poses per person. Five of each pose of persons are used to train HMM, and the remaining 5 are used for testing. Both training and test images go through face recognition process which is divided into seven steps as will be explained in details here. These steps are: filtering, generating observation vectors, feature extraction, feature selection, quantization, training and face recognition. In filtering, a 3×3 minimum filter is applied to the face image, to remove unwanted artifacts, such as highlights in subjects eyes due to flash, and salt noise. After the filtering is done, the face image is converted into one dimensional sequence. It is done by sliding a L×W window from top to bottom of the image, which creates a sequence of overlapping blocks of width W and height L of each face image of width W and height H. Next, the features are extracted. Here, instead of using gray values of pixels in sampling windows, SVD coefficients are used as features. Once the features are extracted a subset of features that lead to smallest classification error and computational cost are extracted from SVD which contains three matrices (**U**, Σ and **V**) two first coefficients of Σ (Σ_{11} and Σ_{22}) and first coefficient of **U** (U_{11}) are used to associate each block. This decreases significantly the length of observation vectors and also computational complexity and sensitivity to noise, changes in illumination, shift and rotation. Since SVD coefficients have innately continuous values, which can lead to infinite number of possible observation vectors, that can't be modeled by discrete HMM, the features need to be quantized. Considering vector $X = (x_1, x_2, ..., x_n)$ the quantized value of x_i is computed as below:

$$x_{iquantized} = \left[\frac{x_i - x_{imin}}{(x_{imax} - x_{imin}) / Q_i} \right] \tag{4.1}$$

x_{imax} and x_{imin} are the maximum and minimum that x_i can get in all possible observation vectors respectively and Q_i is the number of distinct levels to quantize to. Here the first feature (Σ_{11} is quantized into 10, second (Σ_{22} into 7 and third (U_{11}) into 18 levels. After each face image is represented by observation vectors, they are modeled by seven-state HMM. The Baum-Welch algorithm [45] is used to train HMM model for each person in the database. Finally, for each test image the probability of the observation vector X in respect to each HMM face model λ is calculated for classification. A person on a test image m is classified as person d if:

$$P\left(X^{(m)} \mid \lambda_d\right) = \max_n P\left(X^{(m)} \mid \lambda_n\right) \tag{4.2}$$

Figure 4.1 shows an overview of the proposed method using CNN SR method.

4.2 Spare Representation based SR

First method adopted for SR purposes in this thesis was sparse presentation based SR [1]. This method relies on compact representation (for improving performance) of LHR patch pairs sampled from input image. It uses two coupled dictionaries D_h for HR patches, and D_l for LR ones (instead of large training patch database). The sparse representation of LR patch in terms of D_l is directly used to recover the corresponding HR patch from D_h. Patches can overlap if reconstructed high-resolution patches agree on overlapping areas. This method is naturally robust to noise, and therefore the proposed algorithm can handle SR with noisy inputs in a more unified framework.

The basic idea of face image SR consists of two main steps. First, recover a medium high resolution image from face subspace by using reconstruction constraint, which states that LR image I_{LR} is a blurred and downsampled version of the HR image I_{HR}

$$I_{LR} = Sf_b I_{HR} \tag{4.3}$$

where f_b represents a blurring filter and S is the downsampling operator. Second, recover image details using local sparse model - infer patches of HR image for each LR image patch

$$p \approx D_h \alpha \quad \text{for some } \alpha \in \mathrm{R}^K \text{ with } \|\alpha\|_0 \ll K \tag{4.4}$$

The high-resolution image I_{HR} patches p can be represented as sparse linear combination in a HR patches dictionary D_h. The sparse representation α is recovered from patches of input image I_{LR} by representing them with respect to LR dictionary D_l. Dictionaries D_l and D_h are learned from a set of training examples $I_{HR} = \{p_1, p_2, ..., p_{n1}\}$. The dictionaries D_l and D_h are learned so that the sparse representation of LR patches $I^l_{LR} = \{p_1, p_2, ..., p_{n2}\}$ is the same as the sparse representation of corresponding high-resolution patches $I^h_{HR} = \{p_1, p_2, ..., p_{n2}\}$. The learning strategy can be formulated as follows

$$\min_{\{D_h, D_l, Z\}} \|I_{HRc} - D_c Z\|_2^2 + \hat{\lambda} \|Z\|_1 \tag{4.5}$$

where

$$I_{HRc} = \begin{bmatrix} \frac{1}{\sqrt{N}} I^h_{HR} \\ \frac{1}{\sqrt{M}} I^l_{LR} \end{bmatrix} \tag{4.6}$$

$$D_c = \begin{bmatrix} \frac{1}{\sqrt{N}} D_h \\ \frac{1}{\sqrt{M}} D_l \end{bmatrix} \tag{4.7}$$

Z is normalization factor, $\hat{\lambda}$ is sparsity regularization parameter and N and M are the dimensions of high- and low-resolution patches. Additionally first- and second-order derivatives are used as feature for LR patch since the high-frequency components help predicting high-frequency content in HR image. Four one-dimensional filters are used to extract derivatives:

$$\begin{aligned} f_1 = [-1, 0, 1], f_2 = f_1^T, \\ f_3 = [1, 0, -2, 0, 1], f_4 = f_3^T \end{aligned} \tag{4.8}$$

Algorithm 1 FLOW SCHEMA OF PROPOSED SYSTEM USING SPARSE REPRESENTATION SR

```
Input: Low Resolution Image and training dictionary
Output: Face Recognition
for each patch of low resolution image,
    • Compute the mean pixel value of the patch.
    • Solve the optimization problem.
    • Generate the high-resolution patch and put it on high resolution image.'
End for
 Using gradient descent, find the closest image which satisfies the reconstruction constraint
 Using HMM and SVD to find the face recognition rate.
```

In this work, images were super resolved using two different types of dictionaries. One, dictionary was generated of natural images which included images of different plant leaves and flowers, zebra pattern, building sides, veichles, fruits, etc. The other type of dictionaries were generated from only face images. All face images were super resolved using the natural dictionary and face images dictionary. Face images dictionary used for a face database was generated from face images that did not include images from the same database for getting more realistic scenario.

4.3 CNN based SR

Figure **4.1** visualize second method used for SR in this thesis - the deep learning convolutional networks [2]. CNN has been around for decades [46], but recently it has shown an explosive popularity somewhat thanks to its success in image classification [47]. Factors of central importance in this progress are:

- More efficient training implementations on modern powerful GPUs [47]
- Proposed Rectified Linear Unit (ReLU) [48]
- wide range of easily acessible data, such as ImageNet [49], for training models

In order to find super resolved images using CNN, first a bicubic interpolation technique is used for upscaling the image to the desired size. The interpolated image is denoted as $I_{interpolated}$. Our goal is to recover from $I_{interpolated}$ an image I_{SR} which is as similar as possible to the ground truth HR image I_{HR}. We still call $I_{interpolated}$ a LR image for the ease of presentation, although it has the same size as I_{HR}. We wish to learn a mapping F, which conceptually consists of three operations:

- **Patch extraction and representation:** this operation extracts (overlapping) patches from the LR image $I_{interpolated}$ and represents each patch as a high dimensional vector. These vectors comprise a set of feature maps, of which the number equals the dimensionality of the vectors.
- **Non-linear mapping:** this operation nonlinearly maps each high dimensional vector onto another high dimensional vector. Each mapped vector is conceptually

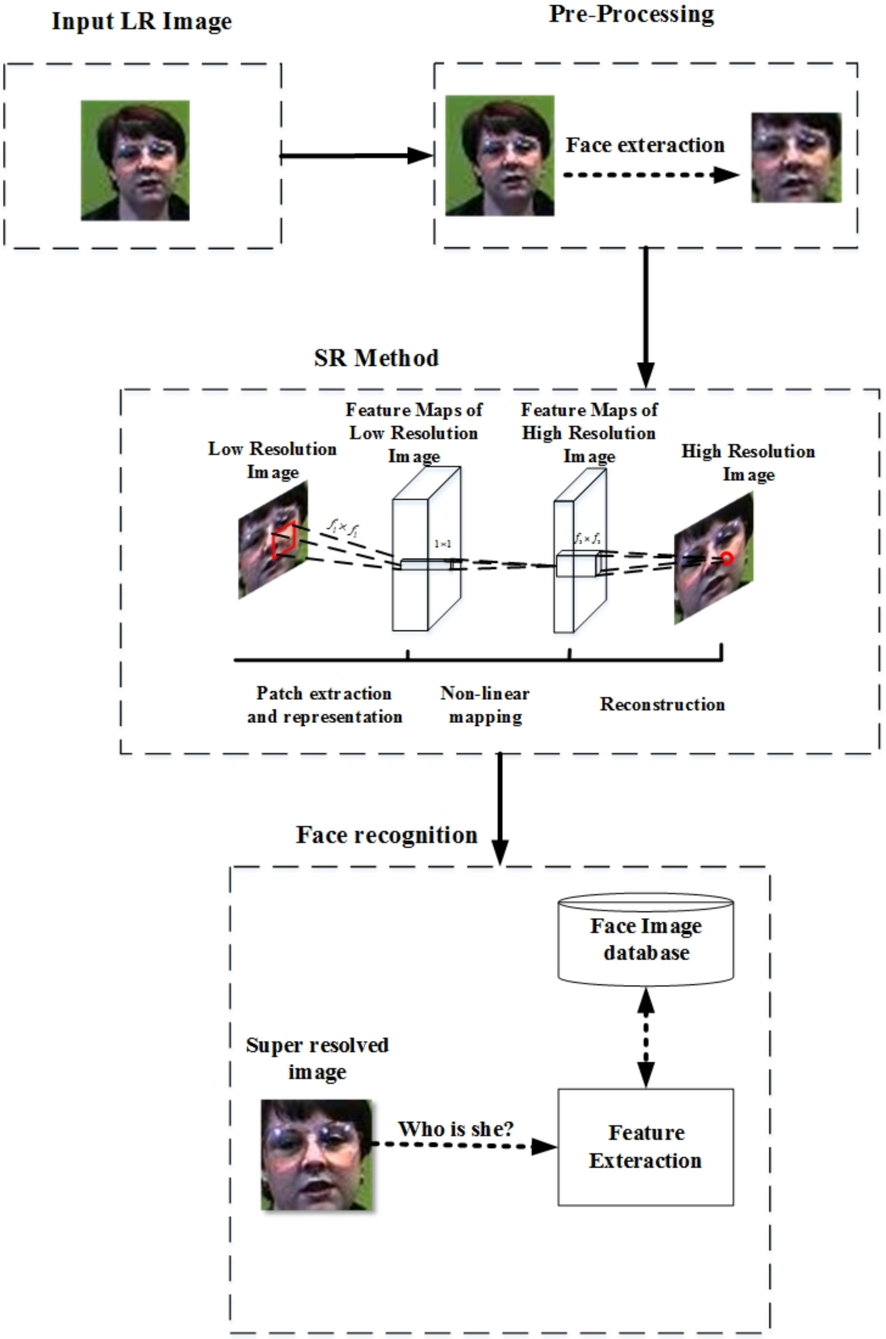

Figure 4.1: The flowchart of the proposed system using CNN SR.

the representation of a HR patch. These vectors comprise another set of feature maps.

- **Reconstruction:** this operation aggregates the above HR patch-wise representations to generate the final HR image. This image is expected to be similar to the ground truth I_{HR}. Figure **4.2** shows an overview of the SR method that is used in the proposed method.

4.3.1 Patch extraction and representation

In image restoration it is common to represent patches that have been densely extracted by a set of pre-trained bases such as PCA, DCT, Haar, etc. Convolving image by a set of filters, each of which is a basis is equivalent to this. In deep learning CNN method these bases are optimized to optimize the network. First layer is expressed as an operation Υ_1:

$$\Upsilon_1\left(I_{interpolated}\right) = \max\left(0, Y_1 * I_{interpolated} + B_1\right) \tag{4.9}$$

where Y_1 represents filters and B_1 biases. The size of Y_1 is $c \times SF_1 \times SF_1 \times n_1$, c represents the number of channels of input image, SF_1 is the spatial size of the filter and n_1 is the number of used filters. Therefore, Y_1 applies n_1 convolutions of size $c \times SF_1 \times SF_1$ on the image. n_1 feature maps are composed for the output. B_1 is a vector that has n_1 dimensions. Each element of B_1 is associated with a filter. ReLU is applied on filter responses.

4.3.2 Non-linear mapping

As a result of first layer, n_1-dimentional feature is extracted for each patch. In second operation, each of these n_1-dimensional vectors are mapped into an n_2-dimensional vector. Second layer operation is:

$$\Upsilon_2\left(I_{interpolated}\right) = \max\left(0, Y_2 * \Upsilon_1\left(I_{interpolated}\right) + B_2\right) \tag{4.10}$$

Y_2 has size of $n_1 \times 1 \times 1 \times n_2$ and B_2 is n_2-dimensional. The output is n_2-dimensional vector, which conceptually represents a HR patch that will be used for reconstruction.

4.3.3 Reconstruction

Predicted overlapping HR patches are often averaged in traditional methods to produce final full image. The averaging can be done by a pre-defined filter on a set of feature maps. Each position in these feature maps is the ”flattened” vector form of a HR patch. Considering this, a convolutional layer to produce the final HR image is defined by:

$$\Upsilon_3\left(I_{interpolated}\right) = Y_3 * \Upsilon_2\left(I_{interpolated}\right) + B_3 \tag{4.11}$$

Here Y_3 size is $n_2 \times SF_3 \times SF_3 \times c$ and B_3 is a vector with c dimensions. If HR patches are represented in the image domain, the filters are expected to behave like averaging filter; if HR patches are represented in some other domains, Y_3 is expected to first project the coefficients onto the image domain and then do the averaging. Y_3 has to be a set of linear filters either way.

4.4 Tracking

Tracking part of this thesis consists of four steps: background extraction, human detection from video frames, face recognition, tracking and presenting information of visited locations in a log file.

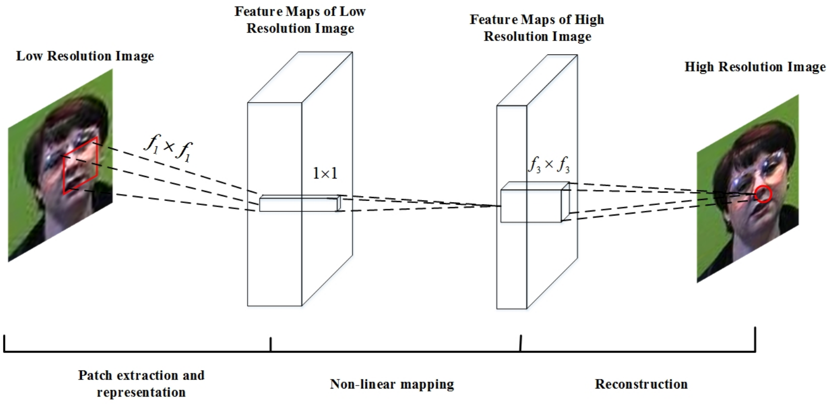
Low Resolution Image
Feature Maps of Low Resolution Image
Feature Maps of High Resolution Image
High Resolution Image
$f_1 \times f_1$
1×1
$f_3 \times f_3$
Patch extraction and representation
Non-linear mapping
Reconstruction

Figure 4.2: The flowchart of the CNN SR method.

4.4.1 Background extraction

Before starting to detect human, the background is extracted from video frame to reduce false detections. There are several ways for removing background from the video frames. We considered mainly two options: average image of frames and using just an empty frame from the video. This tracking system is intended to be used in places like shopping malls, airports, metro, etc. Although these are all crowded areas, there is still a period of time where no people are in the video frame. Such frames can be used as background images. Using average image of multiple frames is both computatively more expensive and has usually shades of foreground items as visible on Figure **4.3**. Considering this, we decided to use single empty frame as a background image I_{bg}. The background extraction from each processed frame is visualized on Figure **4.4** and described in more details here:

- Video frame I_f is subtracted from background to get the difference I_d.

$$I_d = I_{bg} - I_f \tag{4.12}$$

- Difference image is converted into binary image I_{bw}. This function, Υ_b, is shown in equation 4.13. It replaces all pixels with luminance greater than a set level value, l, with 1 and the rest with 0, which correspond to white and black respectively. l is an unsigned integer number in the range from 0 to 255. In this work

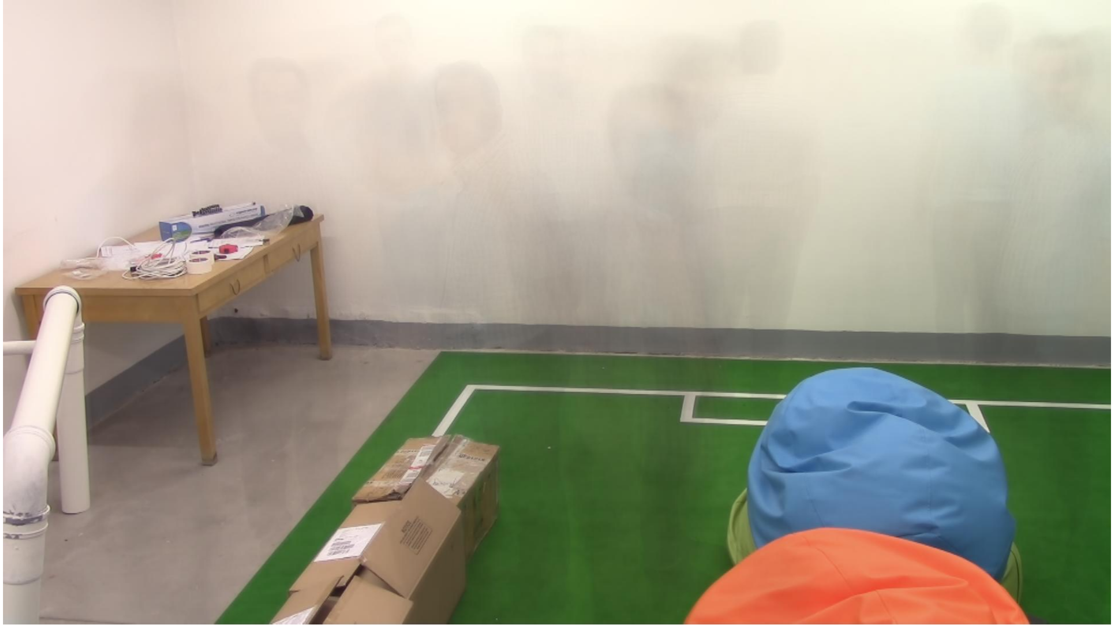

Figure 4.3: Background image from averaging frames with ghost-like noises which are due to moving people in different frames.

we have used mainly 25 as this value.

$$I_{bw} = \Upsilon_b \left(I_d, l \right) \tag{4.13}$$

where Υ_b is an operation to convert an image into the binary image by using the given threshold, l.

- A foreground mask FM is formed from acquired I_{bw}, and frame is then multiplied elementwise by this mask to remove background and get the foreground image I_{fg}.

  $$I_{fg} = FM. \times I_f \tag{4.14}$$

  where $.\times$ denotes element wise operation.

- At this point background pixels of I_{fg} are black - 0 pixel value. These pixels are switched to white for better contrast.

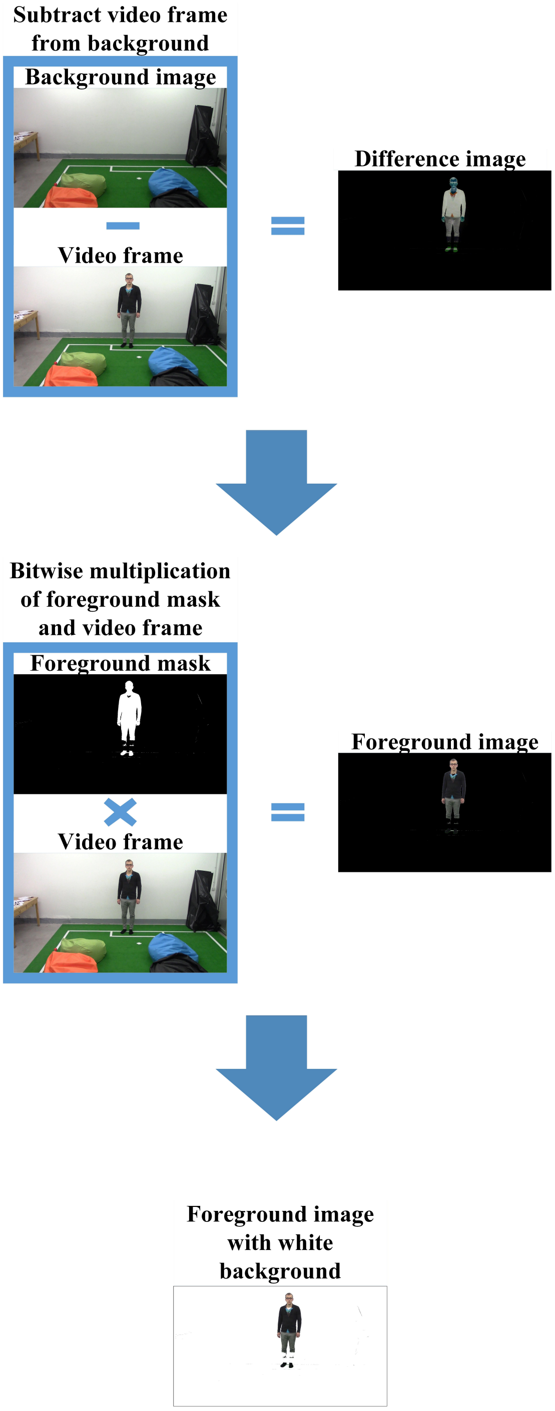
Subtract video frame from background
Background image
Video frame
Difference image
Bitwise multiplication of foreground mask and video frame
Foreground mask
Video frame
Foreground image
Foreground image with white background

Figure 4.4: Flowchart of background extraction.

4.4.2 Human detection

For human detection, we have implemented our own script. First, from foreground image all people and upper bodies are detected. Histogram of Oriented Gradient (HOG) features and a trained Support Vector Machine (SVM) classifier are used to detect people from an input image [50]. Viola-Jones algorithm [5], with specific classification model for upper-bodies is used to detect upper-body regions. Upper-body region is defined as the head and shoulders area. The details of head and shoulder region are encoded using Haar features. Thanks to using more features around the head, this model manages better with pose changes such as head rotations and tilts [51].

A confidence measure is needed for detections for better tracking results. This is handeled by checking if a detected person has also detected upper-body and detected face. If all exist, then the confidence is 1, which is maximum. If two out of three are there, then the confidence is 2/3 and 1/3 in case of only one detection.

When people are moving in crowded areas or between obstructions such as shelves in a store, the most visible part should be the upper-body. Because of this, the detection algorithm is built around upper-body detections. Considering this, for all upper bodies following checks are made:

- For upper-body to be considered together with a detected person, the detected upper-body has to be entirely inside and also entirely above the center of the detected person. Violet horizontal line is visible on figure **4.5**.
- Frontal faces are looked for from inside upper-body. Unlike detecting people and upper bodies, detecting faces is done from original video frame with background. This is done to avoid missing detections due to defects from background extraction. Faces are detected again using Viola-Jones algorithm. But this time with classification model that detects upright and forward facing faces. Model has weak classifiers that use Haar features to encode facial features. Classifiers are composed, based on the classification and regression tree analysis (CART)

which provide the ability to model higher-order dependencies between facial features [52].

  - If no suitable frontal face is found, then another classification model is used to find upright profile faces from inside the upper-body. This model is also composed of weak classifiers, that use Haar features for encoding face details. But unlike for frontal faces, in this model the classifiers are composed based on decision stump.

- If there is either frontal or profile face found from inside the upper-body, then its position relative to the upper-body is checked. If detected face is entirely inside the upper-body and touches the vertical line that goes from the center to the upper edge of the detected upper-body. The line can be seen on figure **4.5**.

4.4.3 Face recognition

In case there is either frontal or frofile face detected in upper-body, face recognition algorithm is ran on it. HMM is again used here. The face is compared against face database. HMM returns index of the closest class and a score. Higher score means that corresponding class is more likely to be best match. For each class in database, there is a threshold score. Score returned by HMM is then compared with the treshold score of returned class. If the score exceeds the threshold, then it is considered to be positive face recognition result. This all is done, because in real world it is almost impossible and in most casses also not necessary to have all people in the database. And if there is a person that is not in a database, then we don't want the system to tell us that this is the person from database that HMM found closest.

4.4.4 Tracking

Once people are detected and face recognition is done, then all this information is saved into a Extensible Markup Language (XML) file. This file contains upper-body detec-

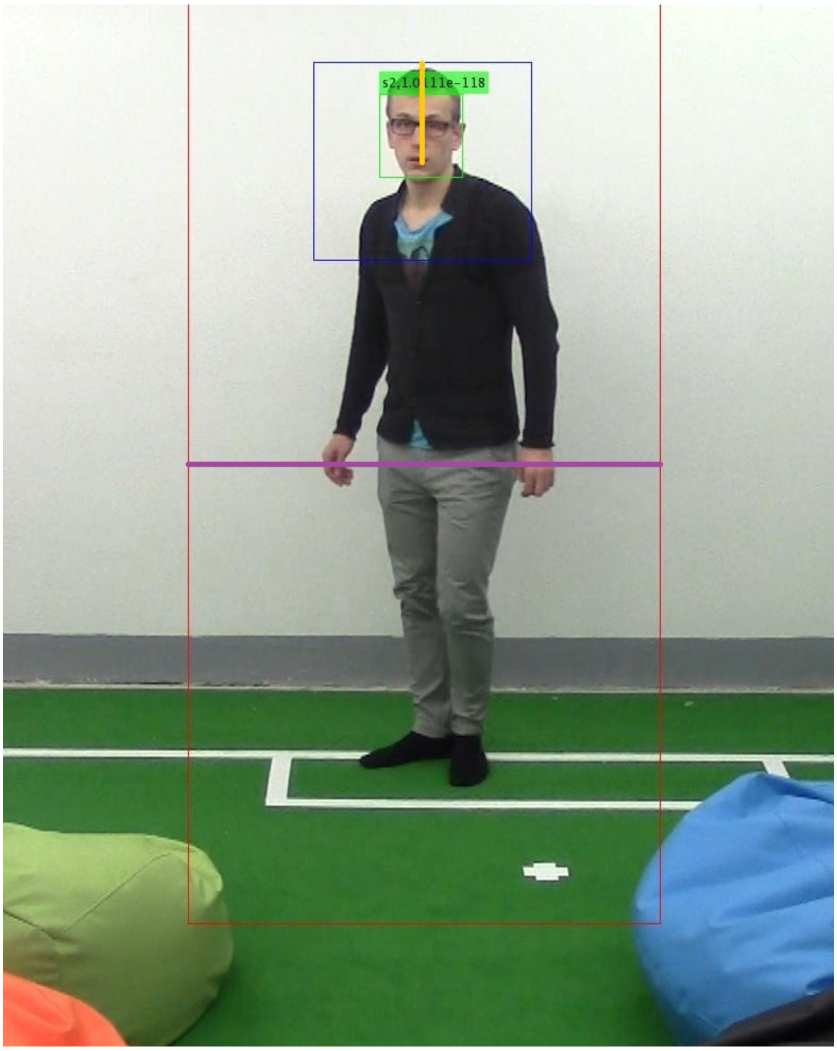

Figure 4.5: Person (red), upper-body (blue) and face (green) detection bounding boxes.

tions, confidence and FR results of each frame. Detection location is presented by height, width, and center coordinates of the detection box among x and y axis. Confidence is presented as floating point number. FR results include: class, score and an indication of whether the score was above threshold or not.

This XML yields as a set of target hypotheses TH for tracking system. In this work we are using a multi-target tracking with a discrete-continuous optimization [6] with some minor modifications. Modifications were made to merge tracks that originated from same detections, for using FR results and to log information of visiting and passing locations.

The goal of the tracking system is to identify a set of target trajectories $T = \{T_1, ..., T_n\}$ from TH. This implies a need for data association a for each detection r and assigning a label $a_r \in \mathbf{LD} = \{1, ..., n\} \cup \emptyset$. As a result, each detection is either identified as belonging to one of the trajectories or has label $\emptyset$, which identifies it as a false alarm. Data association also takes detection confidence into account. Eventually multi-target tracking is performed by minimizing joint energy $E(T$,$a)$ with respect to the T and a.

Once tracking system has produced T, in post processing we check additionally for trajectories that originate from same detections. If there exist two trajectories T_u and T_v, where last n associated detections of T_u are same as first n associated detections of T_v, then T_v is actually an extention of T_u and these trajectories are merged together.

FR result is also assigned to each T. For each trajectory a set of positive FR results (G) is composed. If an associated detection has FR result, that is above threshold, then it is put into G. From this set the most popular FR result is assigned to T.

Another addition was added to tracking system in order to achieve data about locations which tracked people visited or passed. Locations are predefined as bounding boxes. Data about people visiting or passing locations is collected by comparing trajectory coordinates with locations. If trajectory coordinates are inside a bounding box of a location, then it is counted as passing. If trajectory coordinates are inside location for longer than predefined amount of frames, then it is counted as visiting.

5 Experimental results and discussion

All the code for executing experiments is written in MATLAB.

5.1 Experimental results of FR after image SR

Four face databases are used in the experiments. These databases are the facial recognition technology (FERET) database [53, 54], Essex Faces facial images collection [55], Head Pose Image database (HP) [56] and our recently introduced iCV Face Recognition database (iCV-F) [57].

5.1.1 Data

The iCV-F database consists of face images of 31 subjects of which each subject has 10 images. The database includes people wearing glasses or not and various skin color. Models were asked to make different facial expressions while the photos were taken. Fig. 5.1 shows some images of the iCV-F database.

The FERET program was sponsored by the Department of Defense's Counterdrug Technology Development Program through the Defense Advanced Research Products Agency (DARPA). It ran from 1993 through 1997 and its primary mission was to develop an automatic face detection system to assist security, intelligence and law enforcement. FERET database was collected to support testing and evaluation of face

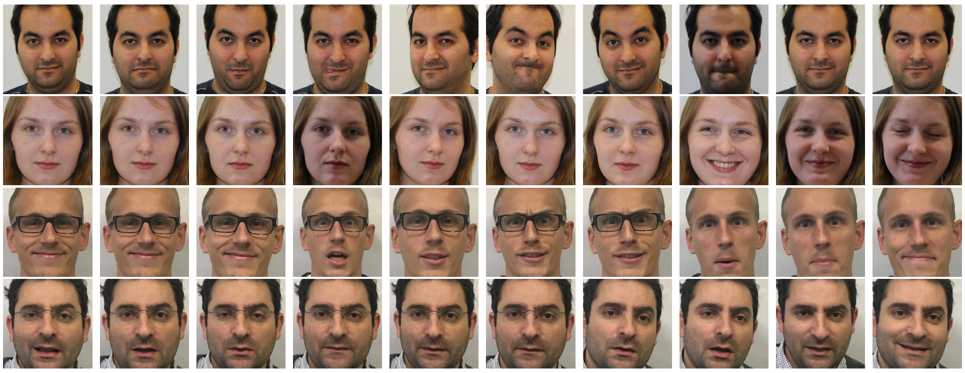

Figure 5.1: Some samples of iCV-F database.

recognition algorithms. The photos were taken in a semi-controlled environment. The same physical setup was used in each photography session to maintain consistency in the whole database. There are some minor differences in images gathered on different dates due to reassembling the equipment for each session. The final corpus consists of 14051 eight-bit grayscale images with views of models' heads ranging from frontal to both left and right profiles. [53,54]. In 2003 a color version of the database was released by DARPA. It included 2413 HR, 24-bit color still facial images of 856 individuals. The selected subset of FERET images database consists of 500 color images.

The Essex Faces consists of 1500 images. The subjects of the database sat at fixed distance from the camera and were asked to speak during the photoshoot. Speaking is for introducing variations of facial expressions. Original images of the database are 180 by 200 pixels. The background of all photos is plain green. No head scale is used. There are very minor changes in head turn, tilt and slant and in position of face in image. No lighting variation in the photos. Also there is no individual hairstyle variation since the photos were taken in a single session [55].

The HP database consists of 2790 face images of 15 individuals with variations of pan and tilt angles from -90 to +90 degrees. It has 2 series of 93 images, all in different poses, for each person. The reason for having 2 series is to have known and unknown faces for training and algorithms. The database features people of various skin color and with or without glasses. The background of images is neutral and uncluttered for focusing on face operations [56]. The subset of the HP database used in this work includes 150 images. Faces in the selected subset are turned only in the horizontal direction.

Our own faces database (iCV-F) consists of 310 images. Photos were taken in 2 sessions in similar conditions. The database includes people wearing glasses or not and various skin color. Models were asked to make different facial expressions while the photos were taken.

5.1.2 Evaluation protocol

The facial images firstly has been passed through Viola-Jones face detector and the segmented faces are resized to 60 x 60 pixels. These images are then downsampled by factor of 4 in order to achieve low resolution input images. Fig. **5.2** shows the LR and super resolved images of different databases. The low resolution images at left have the size of 15x15 and super resolved images at right 60 x 60. The first row is belong to Essex database, the second row is for FERET database, the third and forth rows are belong to HP and iCV-F database respectively.

5.1.3 Results

In order to evaluate and verify the efficiency and reliability of the proposed SR method in terms of providing sufficiently illustrative information for face recognition under various experimentation scenarios, it is applied to numerous databases, where the recognition rates (RR) are obtained for three variants of the images, which are, namely, the original, LR and super-resolved ones. More clearly, the underlying notion is that the performance of the SR technique taken into account is implicitly represented by the capability of the face recognition algorithm in fulfilling its task properly, since it stands for its level of effectiveness in retrieving the data lost at the downsampling stage. The latter

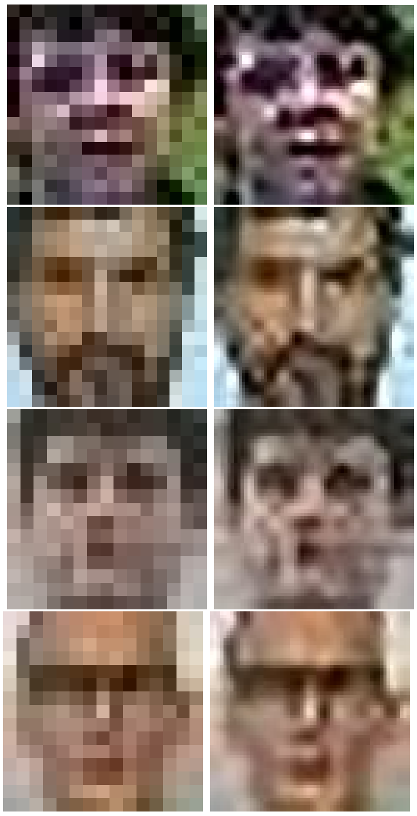

Figure 5.2: LR and SR images of different databases.

is denoted by the amount of improvement appearing in the face recognition rate using the super-resolved images as the subjects, while being compared against that of the LR ones. In other words, when recognizing the faces using the super-resolved images, a more powerful SR method leads to recognition rates closer to the case of considering the original ones.

The databases utilized in the context of the experiments conducted for the purpose of this study include Essex, HP, FERET and iCV-F. Separate recognition rates are reported by using the original, LR and super-resolved images. In order to evaluate the stressfulness of the proposed SR method in enhancing the recognition rate from the case of using the LR images towards the one achieved by taking advantage of the original images, it is shown by the metric compensation ratio (CR), which shows the percentage of the difference between the recognition rate achieved on the LR images and that of the original ones that is compensated by super-resolving the images, being mathematically expressed as follows:

$$CR = \frac{RR_{SR} - RR_{LR}}{RR_O - RR_{LR}} \tag{5.1}$$

where RR_O, RR_{LR} and RR_{SR} stand for the recognition rates accomplished using the original, LR and super resolved images, respectively. For representing the results of sparse representation SR the CR values are separately calculated for natural dictionary (CR_N) and faces dictionary (CR_F) using RR_{SR_N} and RR_{SR_F}, respectively.The results of the above procedure are shown in Tables **5.1** and **5.2**.

Table 5.1: The RR percentages, achieved by sparse representation SR method.

RR	RR_{LR}	RR_{SR_F}	RR_{SR_N}	RR_O	CR_F	CR_N
Essex	75.87	95.07	95.87	99.20	82.30	85.73
iCV-F	76.13	83.87	85.81	98.06	35.29	44.14
HP	26.67	29.33	29.33	50.67	11.08	11.08
FERET	20	14.80	17.60	31.6	-44.83	-20.69

Results of face recognition after super resolving face images using sparse representation based SR method are given in Table **5.1**. This table represents the RR_O, RR_{LR}, RR_{SR_N}, RR_{SR_F} and CR values, in percent, achieved by applying the sparse representation SR method on the Essex, iCV-F, HP and FERET databases, which are sorted in the descending order of CR. Face recognition result for images super resolved using natural dictionary are better than results for faces dictionary for all databases. However if we compare face recognition results of SR images and low resolution image, then we see some differences. CR value of 82.36% and 85.73% for Essex database shows considerable gain in FR results after SR. Also 35.29% and 44.14% for iCV-F are very good compensation rates. CR value for HP database is a bit lower 11.08% and for FERET it is actually negative, which shows that face recognition result for LR images is actually better than for SR images.

All in all, according to results face dictionary doesn't give considerable advantage over natural images dictionary in SR. SR on the other hand gives better results in face recog-

Table 5.2: The RR percentages, achieved by the CNN SR method.

RR	RR_{LR}	RR_{SR}	RR_O	CR
Essex	75.87	86.8	99.20	46.85
HP	26.67	37.33	50.67	44.42
FERET	20	21.60	31.6	13.79
iCV-F	76.13	78.1	98.06	8.98

nition in most cases. Results of face recognition after super resolving face images using CNN based SR method are given in Table **5.2**. This table represents the RR_O, RR_{LR}, RR_{SR} and CR values, in percent, achieved by applying the CNN SR method on the Essex, HP, FERET and iCV-F databases, which are sorted in the descending order of CR. The fact that the CR has taken positive values in the case of all the databases utilized shows that the CNN based SR method has always led to improvements in the face recognition rate, demonstrating its superiority. Using the Essex and HP databases, the latter enhancement is considerable, and takes values as high as 46.85% and 44.42%, respectively. Nevertheless, the recognition rate achieved by using the HP and FERET databases is essentially low, which is due to the lack of robustness in the HMM-based face recognition algorithm [3] against the pose changes, which is widely apparent in both of the foregoing databases. However, the efficiency of the proposed method is clearly shown by the improvements accomplished in recognizing the faces from the Essex and iCV-F databases. Although the recognition rates using the LR images from the foregoing databases are already relatively high, still applying the proposed method has resulted in considerably better recognition rates in both of the cases.

5.1.4 Discussion

In this thesis we have shown that super resolving the low resolution facial images will boost the face recognition rate. Also through experimental result reported in Table 5.1 we have shown that including only facial images in the dictionary will not increase the correct face recognition rate as if we have a dictionary of natural images in the super resolution step. While sparse representation based SR gives FR results close to original images for Essex database, the results for other databases are not so impressive. The CNN based SR images results are more stable giving higher FR results compared to LR images for all databases.

5.2 Experimental results of tracking

5.2.1 Data

In order to test our algorithm for human detection and tracking from video, we recorded a set of videos. Videos were recorded using High Definition camera, hence the resolution was 1920x1080 pixels. Frame rate of the videos was 25 frames per second. The actors in the videos simulated a scenario of customers moving between store shelves when shopping. On hand items like bean bag chairs were used as substitutes to represent store shelves. Actors moved around between shelves and stopped in some locations for 2-3 seconds to represent visiting of a location.

5.2.2 Results

In order to demonstrate the proposed method the results of one video are presented here. This video features three people. The length of the video 1028 frames. In order to reduce the processing time only every fifth frame is processed to find detections. The trajectories of people in the video are presented on a 3D plot on figure **5.3**. x and y-axis on the horizontal plane illustrate the dimensions of each video frame. the vertical axis represents each frame. Therefore, each point on the plot visualize the coordinates of a trajectory on each frame. In total 13 trajectories were found by the tracking algorithm. In ideal there should have been one track for each person. Due to some gaps in detections, the tracking algorithm was unable to merge all trajectories of one person together.

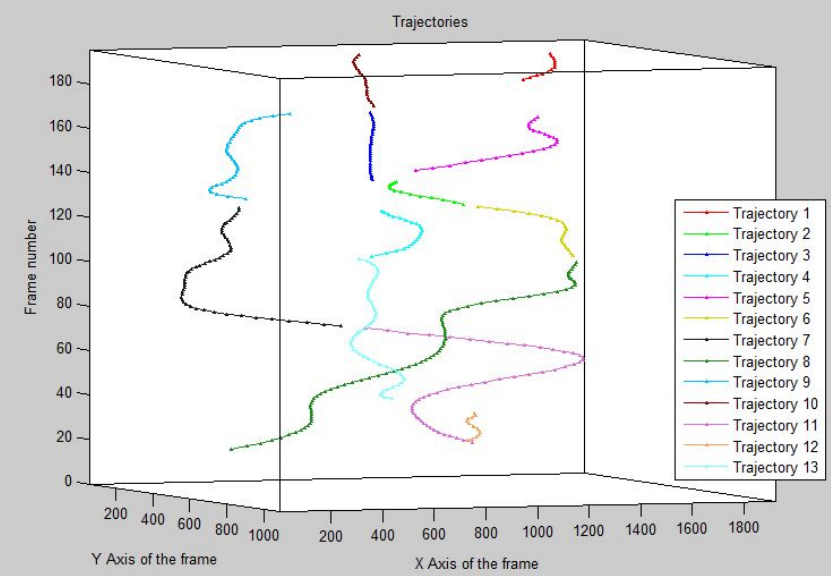

Figure 5.3: Resulting trajectories of tracking

As a result of the tracking process, we know the position of each person by the coordinates of bottom center of upper boby (see trajectory 11 (violet line) on figure **5.4**. The tracking line starts from the point in bottom center of the upper body). These coordinates can also be used to tell if a person is in some location. Therefore, before starting to analyse the movement of people on the video, the locations of interest can be defined. Figure **5.5** visualize the locations defined for our test video. Figure **5.4** shows 2 people at a predefined location: person with trajectory 13 (light blue) stationary in location 7, person with trajectory 8 (green) passing location 4.

According to our scenario, we are interested in the sequence of locations people passed and visited. Table **5.3** represents the order in which each trajectory visited and passed the predefined locations.

In addition to analysing the movement of people, it is possible to also run face recognition on them. In real life scenarios we are not interested in recognizing all people, but only certain individuals of interest that are in the database. To mock this situation, we made a video where two people out of three are known to our face recognition system as classes ’s1’ and ’s2’ and the third person is unknown. For each face processed by

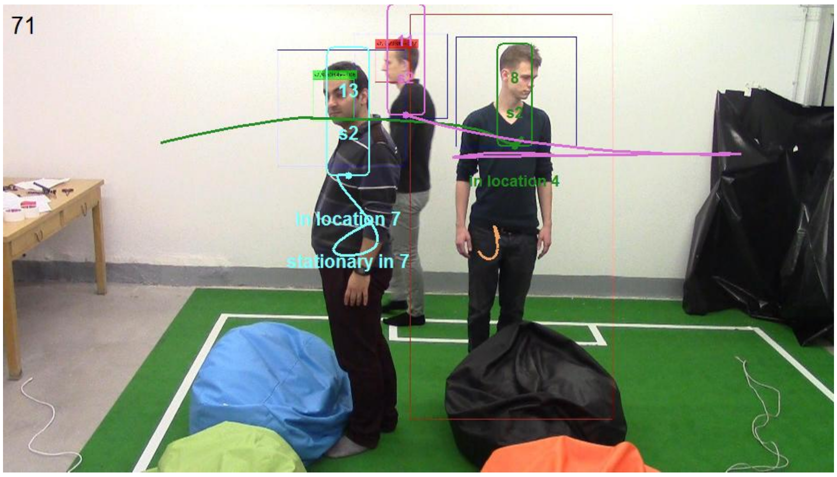

Figure 5.4: Tracked people and indication of visiting or passing locations.

our HMM FR system a class and a score is returned. Class refers to the person in our database that the input face most likely belongs to. Score refers to how similar is the input face to the classified person in the database. Finding the most similar person from the database does not satisfy our aim. We need to determine if the person in the video really is the recognized person. To tell if the FR result is just the closest class or actually the person on the video, a mean face recognition score of each class in the database is found. If a FR result score is higher than the threshold score, then the recognition result is more likely to be actual person in our database, rather than just the closest person. In our experiments we first found the class and score using HMM FR, then compared the score with the threshold score of the class. If score was higher than the threshold, then the FR result was counted as positive, and negative otherwise. Positive meaning here that person on video is actually person from our database. Taking into account all positive FR results for a trajectory, the most popular class is picked as FR result for the trajectory. Table **5.4** represents the results. The FR results returned from the system are compared with actual classes given manually. For the test video the accuracy of correctly classifying people is measured by the percentage of system correctly assigning class to trajectories ant the result was 38.46%. This result is lower than expected and shows that perhaps a more robust FR method should be used here.

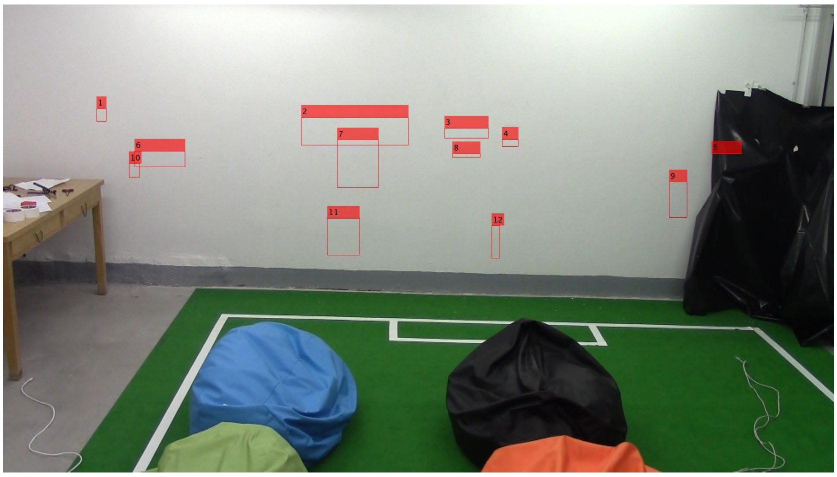

Figure 5.5: Predefined locations

Table 5.3: Visited and passed locations of trajectories

Trajectory	Visited	Passed
1		9
2		4
3	11	11
4	3	2, 3
5		3
6	9	9
7	1, 10	1, 10, 6
8	2, 4	2, 3, 4, 9
9	6	6
10	2	7, 2
11	8	8, 5, 3
12	12	12
13	7, 2	11, 7, 2

Table 5.4: FR results assigned to trajectories

Trajectory	FR result	Actual class
1	s2	s1
2	s2	unknown
3	s2	unknown
4	s2	s1
5	s2	s2
6	s2	unknown
7	s2	s2
8	s2	unknown
9	s2	s1
10	unknown	unknown
11	s2	s2
12	s1	s1
13	s2	s1
Accuracy		38.46%

6 Conclusion and Future Work

This work includes two main parts. First increasing face recognition accuracy from low resolution input by applying super resolution on the facial image before running the FR. Second a method for analysing movement and recognizing people from surveillance videos was proposed.

Two super resolution methods were proposed in first part in order to solve the challenge of face recognition in surveillance videos. First, sparse representation based SR method with a specific dictionary involving facial and natural images. Second, deep learning convolutional network SR method. Face recognition is done after super resolution by adopting HMM and SVD. The system has been tested on many well-known face databases such as FERET, HP, and Essex University databases as well as our own face database. The experimental results shows that the recognition rate is increasing considerably after applying the super resolution.

In the second part a new system for detecting people from video stream was proposed. Detection locations and face recognition results from the detecting system was used as an input for Multi-target tracking system with discrete-continuouos energy minimization. Tracking system was modified to provide information about visiting locations and face recognition results. The system was tested on self made videos.

The results of the second part of this work imply that somewhat better face recognition results could be achieved by using more robust face recognition method in the future. Tracking results could also be slightly improved in the future by improving the detection system by applying more robust method for detecting people. In case of low-resolution input video, image SR can be implemented into detecting system to boost the results.

Acknowledgements

I would like to thank my supervisors Gholamreza Anbarjafari and Pejman Rasti for all the advice and guidance on writing this thesis. Additionally I would like to thank all the people from Institute of Technology of Tartu University who helped us put together our own facial images database by allowing us to take photos of them and also those who helped us with test videos.

References

[1] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861–2873, 2010.

[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," 2015.

[3] H. Miar-Naimi and P. Davari, "A new fast and efficient hmm-based face recognition system using a 7-state hmm along with svd coefficients," *Iranian Journal of Electrical and Electronic Engineering*, vol. 4, no. 1, pp. 46–57, 2008.

[4] F. Lin, C. Fookes, V. Chandran, and S. Sridharan, "Super-resolved faces for improved face recognition from surveillance video," in *Advances in Biometrics*. Springer, 2007, pp. 1–10.

[5] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[6] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1926–1933.

[7] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*. Springer, 2012, pp. 711–730.

[8] J. Yang, Z. Lin, and S. Cohen, "Fast image super-resolution based on in-place example regression," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 1059–1066.

[9] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *Image Processing, IEEE Transactions on*, vol. 23, no. 6, pp. 2569–2582, 2014.

[10] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine vision and applications*, vol. 25, no. 6, pp. 1423–1468, 2014.

[11] P. Rasti, H. Demirel, and G. Anbarjafari, "Image resolution enhancement by using interpolation followed by iterative back projection," in *Signal Processing and Communications Applications Conference (SIU), 2013 21st*. IEEE, 2013, pp. 1–4.

[12] P. Rasti, I. Lusi, A. Sahakyan, A. Traumann, A. Bolotnikova, M. Daneshmand, R. Kiefer, A. Aabloo, G. Anbarjafar, H. Demirel *et al.*, "Modified back projection kernel based image super resolution," in *Artificial Intelligence, Modelling and Simulation (AIMS), 2014 2nd International Conference on*. IEEE, 2014, pp. 161–165.

[13] L. Wang, S. Xiang, G. Meng, H. Wu, and C. Pan, "Edge-directed single-image super-resolution via adaptive gradient magnitude self-interpolation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 8, pp. 1289–1299, 2013.

[14] E. Mjolsness, "Fingerprint hallucination," Ph.D. dissertation, California Institute of Technology, 1985.

[15] S. Baker and T. Kanade, "Hallucinating faces," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 2000, pp. 83–88.

[16] ——, "Super-resolution: Reconstruction or recognition," in *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing*, vol. 153, 2001.

[17] ——, "Limits on super-resolution and how to break them," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 9, pp. 1167–1183, 2002.

[18] ——, "Super-resolution: Limits and beyond," in *Super-Resolution Imaging*. Springer, 2002, pp. 243–276.

[19] W. T. Freeman and E. Pasztor, "Markov networks for low-level vision," *Mitsubishi Electric Research Laboratory Technical, Report TR99-08*, 1999.

[20] A. J. Storkey, "Dynamic structure super-resolution," in *Advances in Neural Information Processing Systems*, 2002, pp. 1295–1302.

[21] D. Capel and A. Zisserman, "Super-resolution from multiple views using learnt image models," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2. IEEE, 2001, pp. II–627.

[22] H. Yan, J. Liu, J. Sun, and X. Sun, "Ica based super-resolution face hallucination and recognition," in *Advances in Neural Networks–ISNN 2007*. Springer, 2007, pp. 1065–1071.

[23] Y. Liang, J.-H. Lai, X. Xie, and W. Liu, "Face hallucination under an image decomposition perspective," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 2158–2161.

[24] F. M. Candocia and J. C. Principe, "Super-resolution of images based on local correlations," *Neural Networks, IEEE Transactions on*, vol. 10, no. 2, pp. 372–380, 1999.

[25] D. Lin, W. Liu, and X. Tang, "Layered local prediction network with dynamic learning for face super-resolution," in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 1. IEEE, 2005, pp. I–885.

[26] A. J. Tatem, H. G. Lewis, P. M. Atkinson, and M. S. Nixon, “Super-resolution target identification from remotely sensed images using a hopfield neural network,” *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, no. 4, pp. 781–796, 2001.

[27] C. Miravet and F. B. Rodríguez, “A hybrid mlp-pnn architecture for fast image superresolution,” in *Artificial Neural Networks and Neural Information Processing-ICANN/ICONIP 2003*. Springer, 2003, pp. 417–424.

[28] E. Salari and S. Zhang, “Integrated recurrent neural network for image resolution enhancement from multiple image frames,” in *Vision, Image and Signal Processing, IEE Proceedings-*, vol. 150, no. 5. IET, 2003, pp. 299–305.

[29] C. Miravet and F. B. Rodríguez, “Accurate and robust image superresolution by neural processing of local image representations,” in *Artificial Neural Networks: Biological Inspirations–ICANN 2005*. Springer, 2005, pp. 499–505.

[30] B. V. Kumar and R. Aravind, “Face hallucination using olpp and kernel ridge regression,” in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*. IEEE, 2008, pp. 353–356.

[31] K. Jia and S. Gong, “Multi-modal face image super-resolutions in tensor space,” in *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*. IEEE, 2005, pp. 264–269.

[32] M. Turk, A. P. Pentland *et al.*, “Face recognition using eigenfaces,” in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR’91., IEEE Computer Society Conference on*. IEEE, 1991, pp. 586–591.

[33] A. Pentland, B. Moghaddam, and T. Starner, “View-based and modular eigenspaces for face recognition,” in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR’94., 1994 IEEE Computer Society Conference on*. IEEE, 1994, pp. 84–91.

[34] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 711–720, 1997.

[35] W. Zhao, R. Chellappa, and N. Nandhakumar, “Empirical performance analysis of linear discriminant classifiers,” in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on.* IEEE, 1998, pp. 164–169.

[36] G. Guo, S. Z. Li, and K. Chan, “Face recognition by support vector machines,” in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on.* IEEE, 2000, pp. 196–201.

[37] F. S. Samaria and A. C. Harter, “Parameterisation of a stochastic model for human face identification,” in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on.* IEEE, 1994, pp. 138–142.

[38] V. V. Kohir and U. B. Desai, “Face recognition using a dct-hmm approach,” in *Applications of Computer Vision, 1998. WACV’98. Proceedings., Fourth IEEE Workshop on.* IEEE, 1998, pp. 226–231.

[39] F. S. Samaria, “Face recognition using hidden markov models,” Ph.D. dissertation, University of Cambridge, 1994.

[40] S. Eickeler, S. Müller, and G. Rigoll, “Recognition of jpeg compressed face images based on statistical methods,” *Image and Vision Computing*, vol. 18, no. 4, pp. 279–287, 2000.

[41] C. Anand and R. Lawrance, “Algorithm for face recognition using hmm and svd coefficients,” *Artificial Intelligent Systems and Machine Learning*, vol. 5, no. 3, pp. 125–130, 2013.

[42] M. Bicego, U. Castellani, and V. Murino, “Using hidden markov models and wavelets for face recognition,” in *Image Analysis and Processing, 2003. Proceedings. 12th International Conference on.* IEEE, 2003, pp. 52–56.

[43] J. Bobulski, “2dhmm-based face recognition method,” in *Image Processing and Communications Challenges 7*. Springer, 2016, pp. 11–18.

[44] V. C. Klema and A. J. Laub, “The singular value decomposition: Its computation and some applications,” *Automatic Control, IEEE Transactions on*, vol. 25, no. 2, pp. 164–176, 1980.

[45] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[46] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[48] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[50] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[51] H. Kruppa, M. Castrillon-Santana, and B. Schiele, “Fast and robust face finding via local context,” in *Joint IEEE Internacional Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, 2003, pp. 157–164.

[52] R. Lienhart, A. Kuranov, and V. Pisarevsky, “Empirical analysis of detection cascades of boosted classifiers for rapid object detection,” in *Pattern Recognition*. Springer, 2003, pp. 297–304.

[53] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, “The feret database and evaluation procedure for face-recognition algorithms,” *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[54] P. J. Phillips, H. Moon, S. Rizvi, P. J. Rauss *et al.*, “The feret evaluation methodology for face-recognition algorithms,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 10, pp. 1090–1104, 2000.

[55] Collection of facial images: Faces94. [Online]. Available: http://cswww.essex.ac.uk/mv/allfaces/faces94.html

[56] Head pose image database. [Online]. Available: http://www-prima.inrialpes.fr/perso/Gourier/Faces/HPDatabase.html

[57] Collection of facial images. [Online]. Available: http://icv.tuit.ut.ee/databases.html

Appendix A: Publications

This thesis has produced two conference papers:

- Tõnis Uiboupin, Pejman Rasti, Gholamreza Anbarjafari, Hasan Demirel, ”Facial Image Super Resolution Using Sparse Representation for Improving Face Recognition in Surveillance Monitoring”, IEEE SIU2016
- Pejman Rasti, Tõnis Uiboupin, Sergio Escalera and Gholamreza Anbarjafari, ”Convolutional Neural Network Super Resolution for Face Recognition in Surveillance Monitoring”, AMDO 2016

Appendix B: Source codes

All code for testing proposed methods in this thesis were written in MATLAB. Produced code is available in Github at https://github.com/tuiboupin/SRFRBPAMEUSC

Non-exclusive licence to reproduce thesis and make thesis public

I, Tõnis Uiboupin (date of birth: 7th of October 1988),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

RESOLUTION AND FACE RECOGNITION BASED PEOPLE ACTIVITY MONITORING ENHANCEMENT USING SURVEILLANCE CAMERA

supervised by Gholamreza Anbarjafari and Pejman Rasti

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu 28.03.2016