2008

Adaptive frame selection for enhanced face recognition in low-resolution videos

Raghavender Reddy Jillela
West Virginia University

Adaptive Frame Selection for Enhanced Face Recognition in Low-Resolution Videos

by

Raghavender Reddy Jillela

Thesis submitted to the
College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Electrical Engineering

Arun Ross, PhD., Chair
Xin Li, PhD.
Donald Adjeroh, PhD.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia
2008

Keywords: Face Biometrics, Super-Resolution, Optical Flow, Super-Resolution using Optical Flow, Adaptive Frame Selection, Inter-Frame Motion Parameter, Image Quality, Image-Level Fusion, Score-Level Fusion

Abstract

Adaptive Frame Selection for
Enhanced Face Recognition in Low-Resolution Videos

by

Raghavender Reddy Jillela
Master of Science in Electrical Engineering

West Virginia University

Arun Ross, PhD., Chair

Performing face detection and recognition in low-resolution videos (e.g., surveillance videos) is a challenging task. To enhance the biometric content in these videos, image-level and score-level fusion techniques can be used to consolidate the information available in successive low-resolution frames. In particular, super-resolution can be used to perform image-level fusion while the simple sum-rule can be used to perform score-level fusion. In this thesis we propose a technique which adaptively selects low-resolution frames for fusion based on optical flow information. The proposed technique automatically disregards frames that may cause severe artifacts in the super-resolved output by examining the optical flow matrices pertaining to successive frames. Experimental results demonstrate an improvement in the identification performance when adaptive frame selection is used to perform super-resolution. In addition, improvements in output image quality and computation time are observed. In score-level fusion, the low-resolution frames are first spatially interpolated and the simple sum rule is used to consolidate the match scores generated using the interpolated frames. On comparing the two fusion methods, it is observed that score-level fusion outperforms image-level fusion. This work highlights the importance of adaptive frame selection in the context of fusion.

I dedicate my thesis to my parents

Acknowledgments

This thesis has been made possible by many. Firstly, I would like to thank Dr. Arun Ross for being my advisor and the committee chair. His dedication, patience and hard-working nature have inspired me in many ways. Working with him has taught me to conduct research with rigour, and to present my work with clarity. I consider it as a privilege to have him as my advisor and instructor throughout my Master's program.

I would also like to thank Dr. Xin Li and Dr. Donald Adjeroh, my committee members, for their valuable guidance and suggestions.

I would like to extend my gratitude to all my lab mates Aglika, Asem, Ayman, Chen, David, Ding, Manisha, Matt, Matthias, Nick, Nikhil, Phani, Raghu, Rajiv, Simona and Susan for their constructive criticism and camaraderie in and out of work.

I thank all my friends and relatives who always believed in me.

I will forever be indebted to my family for their unconditional love, constant support and motivation.

Finally, I thank God for everything that is memorable in my life.

Contents

List of Figures

Chapter 1

Introduction

1.1 Face as a Biometric

The science of establishing the identity of an individual using physical or behavioural traits by employing automated or semi-automated techniques is known as biometrics. The primary tasks of a biometric system include verification, identification and surveillance. Some of the physical and behavioural traits of an individual that can be used for authentication purposes include: fingerprint, face, iris, hand geometry, DNA, retina, signature, voice, palmprint, gait, hand vein pattern, saccadic movements, ear, key stroke dynamics, and facial and hand thermograms.

Among the above mentioned list of biometric traits, face has a very high significance in human authentication because of the following properties:

- Universality: Face is one of the most common biometric traits possessed by all humans.
- Collectability: Facial information is very easily collectable, requiring minimal user interaction with sensors compared to other biometric systems.
- Acceptability: Face is widely considered as one of the most non-intrusive biometric feature to acquire.

Because of these properties, face biometrics has been extensively researched. Some of the research areas related to face biometrics are [5]:

(a) Face detection: The goal of face detection is to determine whether or not there are any faces in a given arbitrary image, and if present, to determine the location and extent of each face.

(b) Face tracking: The process of continuously estimating the location and possibly the orientation of a face in a image sequence in real time is termed as face tracking.

(c) Facial feature detection: A facial feature detection system detects the presence and location of features such as eyes, nose, nostrils, eyebrow, mouth, lips, ears, etc. with the assumption that there is only one face in an image.

(d) Face identification: A face recognition system compares an input image (probe) against multiple gallery images in order to determine the identity of the probe.

(e) Face verification: The purpose of a face authentication system is to verify the claim of the identity of an individual using an input face image.

Face detection and face recognition (both verification and identification) are the most widely researched areas in the face biometric domain. Many commercial applications have been developed based on these techniques and have been employed in real world problems such as airport security, access control, surveillance, and smart environments at home and in cars [6], [7]. Though face has been considered as one of the most easily collectable biometric feature and face biometric systems exhibit good matching performance, many challenges still exist, providing a scope for improvement in many ways [8]. Some of the most significant challenges associated with face detection or face recognition systems are listed below:

(1) Variations in pose: Depending on the relative camera-face pose, images of the face of the same individual might vary substantially. This might also occlude some of the facial features making the detection and recognition tasks difficult. Figure 1.1 shows the left, frontal and right profiles of an individual, illustrating the changes in appearance due to pose variations.

(2) Illumination changes: The appearance of a face can be hugely impacted by the characteristics of illuminating source such as source distribution and intensity.

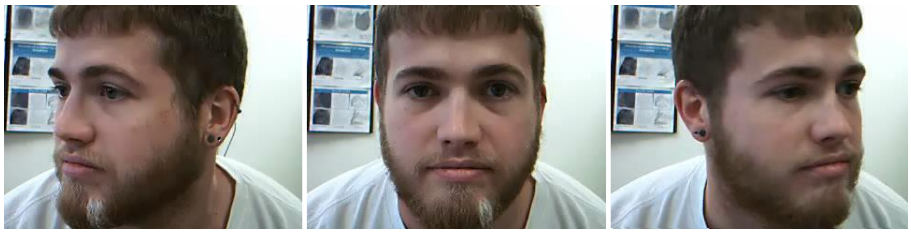

Figure 1.1: Left, frontal, and right profiles of an individual.

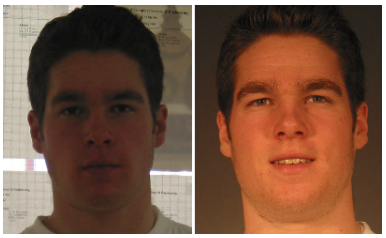

Figure 1.2: Example of illumination variation [1].

(3) Variation in expressions: The appearance of a face is dependent on the expression of the individual. As seen in Figure 1.3, expressions such as smiling, laughing and neutral face can change the apperance of an individual.

Figure 1.3: Example of expression variation [2].

(4) Occlusions: Face biometric systems detect or recognize faces by extracting features from images. In cases where faces are partially occluded by objects, extracting some of the features would be impossible which reduces the matching accuracy of the system. Example of partially occluded faces are shown in Figure 1.4.

(5) Image orientation: Appearance of a face in an image varies with the different angles of rotation of face about the camera's optical axis. This makes the task of detection and recognition of an individual very difficult. Figure 1.5 shows an individual's facial image captured with different angles of rotation about the camera's optical axis.

Figure 1.4: Example of partially occluded faces.

Figure 1.5: Facial images of an individual with different orientations.

(6) Aging: With increasing age, the appearance of human face changes. This includes variations in size and shape of the face, and possibilities of variations in skin color and texture. An example of variation in the face with age is shown in Figure 1.6.

Figure 1.6: Variation in appearance of an individual due to aging [3].

(7) Presence of structural components: Structural components such as beards, moustaches, glasses, or hats can alter the appearance of an individual by a great extent. Figure 1.7 illustrates variations in the appearance of an individual due to the presence of structural components.

Like several other biometric features, face biometrics can be used for verification, identification or surveillance. These tasks can be carried out by using either images or videos containing facial images in static or dynamic backgrounds. Some of the applications using face as a biometric include access control, law enforcement and surveillance tasks. Access control usually occurs in a controlled setting where individuals voluntarily submit their facial profile. In surveillance tasks, individuals are not expected to cooperatively

Figure 1.7: Appearance variations due to presence of structural components [2].

interact with cameras, thereby resulting in an uncontrolled setting. Law enforcement tasks usually deal with both scenarios ranging from static mug-shot verification to identification in a dynamic background.

1.2 Face Biometrics in Surveillance Applications

1.2.1 Surveillance

Surveillance, in a broad sense, can be defined as the monitoring of activities of individuals or groups from a position of higher authority. Surveillance serves two main needs of a society: security, and safety in life and work activities. Surveillance can be carried out in different domains ranging from traffic regulation to monitoring places of interest, or to merely deter individuals from committing unlawful activities. Surveillance can be carried out either as:

- a *covert* operation, in which individuals do not have any knowledge about surveillance activities, or as
- an *overt* operation, where individuals know that they are subject to some kind of monitoring, and are provided with reminders of monitoring actions.

1.2.2 Biometric Surveillance

Biometric surveillance is a subset of surveillance where individuals in consideration can be monitored by automated systems dependent on biometric features. With an increasing demand for security and safety, biometric surveillance is becoming an important technology in today's world. Face is one of the most widely used biometric for biometric surveillance. Although other biometric features like the iris could potentially be used

for surveillance purposes, usage of face biometric systems is more widespread due to its properties of universality, acceptability and collectability. Biometric surveillance using facial features is carried out prominently by recording the images or videos of individuals. Surveillance images and videos captured using various sensors are main sources for documenting and archiving the monitored activities of concern.

Surveillance videos are video sequences recorded by visual surveillance systems which are used for monitoring and registering the activities of people, objects or processes in spaces of special interest for security and control purposes. Video surveillance systems are prominently used in security and traffic monitoring. A set of sample frames from surveillance videos recording humans can be seen in Figure 1.8[1].

Figure 1.8: Sample frames from surveillance videos.

1.3 Motivation

Human surveillance videos can be used to identify and track individuals of interest or generate watch lists, depending on the level of automation. Systems can be designed to detect and/or recognize faces in a complex situation, either in an online or offline mode, depending on the level of resources and technology used. In most applications, the task of identifying individuals in surveillance videos is performed by human experts rather than employing automated computer vision routines due to the complexity associated with it.

Though automated face detection or recognition systems for surveillance videos are already in use, there are many concerns which still need to be addressed. Some of the major challenges related to face detection or recognition in surveillance videos are listed below:

(a) Surveillance videos typically involve uncontrolled situations, both with respect to the individual as well as the external conditions. In cases where video surveillance is

[1]http://www.nytimes.com/imagepages/2005/07/22/international/22cnd-london.1.html

carried out as a covert task, any kind of co-operation from the humans in the videos is not expected. Also since the cameras are deployed in an uncontrolled environment, there is a possibility that external conditions, like weather, may cause problems in extracting regions of interest. A good example of the above mentioned case is a surveillance camera employed in a parking lot.

(b) Since surveillance cameras are usually deployed in public spaces such as metro stations, airports, supermarkets, office buildings, hospitals, etc., the ensuing videos record human subjects from a significant distance. Due to this reason, the size of faces in a surveillance video can be very small. This reduces the *inter-pupillary distance* which is a crucial factor in detecting or recognizing faces.

(c) As surveillance videos are captured in an uncontrolled environment, motion blur can be present in the frames of the videos. Motion blur further aggravates the complexity of the human identification problem in surveillance videos.

(d) Adding to these factors, since most surveillance videos are generally low resolution videos, human identification tasks become extremely difficult.

With the increasing demand for face biometric systems, the need to develop better automated systems is growing. Many attempts have been made to improve the performance of biometric systems in surveillance applications [9], [10], [11], [12]. To improve the performance of a face biometric system, focus is placed mainly on the task of retrieving a relatively good facial profile from the videos.

1.4 Problem Statement

Given a low resolution video $\overline{V}$, containing n frames (i.e., images) $\{f_1, f_2, f_3, ..., f_n\}$, it is desired to obtain a subset of frames $\{h_1, h_2, h_3, ..., h_k\}$, $k \leq n$, which have favorable facial information for reliable face detection and recognition. This can be achieved by implementing various video and image processing techniques, to first extract the facial region in the video and then increase the facial information content in the video. Also, it is desired that the images generated by these processes have a better image quality.

Further, it is desired to develop a technique to drop frames from the original sequence based on some inherent criteria of the video so as to improve the recognition performance

using the remaining frames. The technique could drop frames based on any effective criteria such as poor quality, motion blur, pose changes or a combination of these. This would help in effective recognition in cases when it is known that one individual is present in a video and the best quality image has to be retrieved. This technique can also help in an online environment when only good quality images of individuals need to be archived.

1.5 Approach of the Thesis

To achieve the goal of generating a better quality image from a given surveillance video, super-resolution techniques can be applied for increasing the information content in an image. Super resolution techniques process a single frame or a set of frames and output a higher resolution image. This thesis focuses on one technique in particular, super-resolution optical flow, due to its applicability to videos containing human faces.

To drop the frames in a given video, emphasis is laid mainly on the quality aspect of an image. It is attempted to mathematically formulate *motion*, an inherent component of video, which can be used as a liminal to eliminate motion blurred images.

Fusion of frames can be considered as an effective means of obtaining a better image from a given set of images. This concept was described in [13] as a noise removal process, where motion was not present between consecutive frames. This thesis tries to extend the concept by adaptively selecting frames, based on the motion between successive frames, and fusing them for obtaining a good quality output. The frames which are dropped due to the high motion component can be considered as those frames which refer to changes in scene or pose of an individual in the video.

The thesis is organized as follows: first, an introduction to super-resolution and other image processing techniques which can be used to extract maximum facial information from an image, are discussed. In Chapter 3, super-resolution optical flow and the proposed adaptive frame selection technique are explained. Image quality metrics are discussed in chapter four. Results of the experiments conducted and the inferences are included in chapter five. The final chapter discusses the contributions of the thesis and directions for future work.

Chapter 2

Super Resolution

2.1 Introduction

2.1.1 Image Resolution

According to [14], image resolution can be defined as the smallest discernible or measurable detail in a visual presentation. In digital image processing and computer vision, the term resolution can be used in the following ways [14]:

- Spatial resolution, which refers to the spacing of pixels in an image. Usually spacial resolution is measured in pixels per inch (ppi). For a given image, higher the spatial resolution, greater the number of pixels, and correspondingly smaller the size of individual pixels in that image. Higher spatial resolution allows for more detail and subtle color transitions in an image.

- Brightness resolution, usually refers to the number of brightness levels that can be recorded at any given pixel. The brightness resolution for monochrome images is usually 256, implying that one level is represented by 8 bits. For full color images, at least 24 bits are used to represent one brightness level, i.e., 8 bits per color plane (red, green and blue).

- Temporal resolution, which refers to the number of frames captured per second also commonly known as the frame rate. It is related to the amount of perceptible motion between the frames. Higher frame rates result in more clarity and less smearing due

to movements in the scene. Frame rates of 25 frames per second, or above, are usually considered suitable for a pleasing view.

In this work, the term resolution refers to the spatial resolution.

2.1.2 Image Super-Resolution

Super-resolution is the process of generating a raster image of a scene with a higher resolution than its source [15]. A super resolved image possesses higher pixel density when compared to its source image and thus offers more details about the scene captured in the image. Figure 2.1 shows a sample high resolution (HR) image that can be obtained by super resolving a low resolution image(LR).

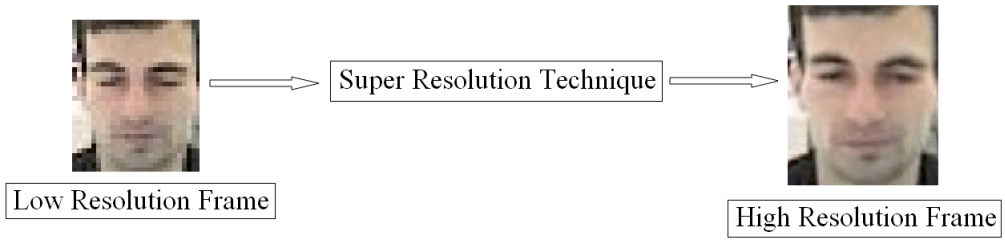

Figure 2.1: Super resolved HR image from a LR image.

The source used to generate a super-resolved image can comprise of a single image or a set of images. Based on the number of source images used, super-resolution techniques may be classified into two groups:

(a) Techniques that use a single image, which interpolate the pixel information available in the image.

(b) Techniques that use multiple images, and output a higher resolution image by fusing the pixel information from multiple observations of the same scene.

A detailed discussion of the various super-resolution techniques, along with their classifications, is provided in the following sections of this chapter.

2.2 Observation Model

To understand and analyze the process of generating HR images from LR images using super-resolution, it is necessary to formulate an observation model. Most of the observation models for super-resolution algorithms are based on a fundamental principle that the super-resolved image when appropriately warped and down-sampled, should generate a set of low resolution input images [4].

Consider a HR image x of size $L_1N_1 \times L_2N_2$, with the parameters L_1 and L_2 representing the down-sampling factors in the observation model along the horizontal and vertical directions, respectively. Each observed LR image of size $N_1 \times N_2$ is represented by y_k where $k = 1, 2, ..., p$ and p is the total number of LR images. It is assumed that x remains constant during the acquisition of the multiple LR images, except for any motion and degradation allowed by the model. Therefore, the observed LR images result from warping, blurring and subsampling operators performed on the HR image $\mathbf{x}$. With the assumption that each LR image is corrupted by additive noise, the observation model can be represented by the following equation:

$$y_k = DB_kM_kx + n_k \text{ for } 1 \leq k \leq p \tag{2.1}$$

where M_k is the warp matrix, B_k represents a blur matrix, D is a subsampling matrix and n_k represents the additive noise that corrupts the image. Figure 2.2 illustrates the described observational model.

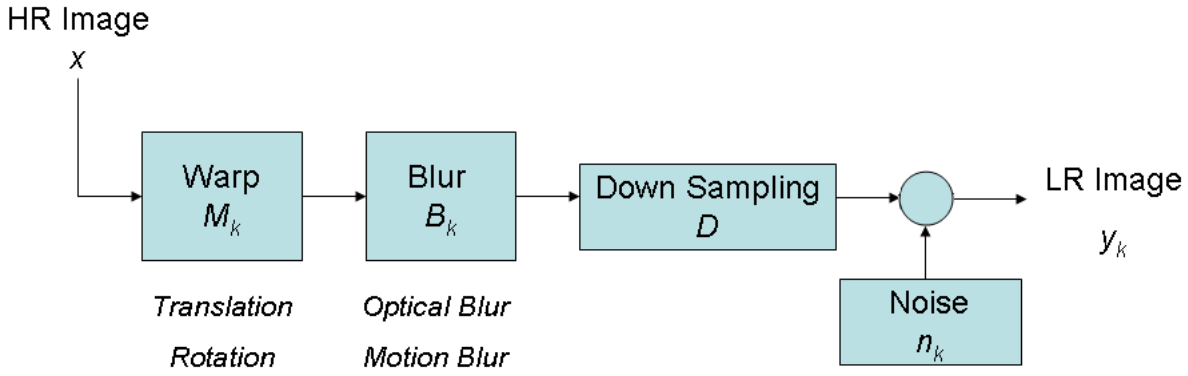

Figure 2.2: Observation model relating LR images to HR images, based on [4].

2.3 Super Resolution Techniques

Based on the algorithm used, super-resolution techniques can be classified into two categories: *Reconstruction - based* and *Recognition - based* [16]. Reconstruction based techniques operate directly on image pixel intensities and super-resolve an image sequence by fusing the information from various observations of the same scene [17]. Recognition based techniques on the other hand, learn the features of low resolution input images and synthesize the corresponding high resolution output [4].

On the other hand, super-resolution techniques can also be broadly divided into techniques for still images and for video sequences, but it can be considered that a video sequence technique is a straight forward extension of still image models [18]. Description of each technique, along with the literature review are presented in the following sections.

2.3.1 Super-resolution using Single Image

Resolution of an image can be increased by interpolating the pixel intensities. A number of interpolation algorithms have been proposed in the image processing domain, ranging from nearest-neighbor, bilinear, to cubic spline interpolation [19], [20].

More sophisticated algorithms, like edge-preserving techniques [21], regularization-based approaches [22], and Bayesian algorithms [23] have been explained in the survey paper of Schultz and Stevenson [24]. Interpolation of an image by modeling of Markov networks is explained in the paper by Freeman and Pasztor [25]. It was noted in the work by Baker and Kanade [26] that while interpolation can give good results when the input images are of fairly high resolution, it often performs worse as the input images get smaller. Also, it was mentioned by Park et. al. [27] that, typically, image interpolation methods cannot be viewed as super-resolution techniques as they cannot recover the high-frequency components of low resolution images.

2.3.2 Super-resolution using Multiple Images

Elad and Feuer [28] described the process of obtaining a high-resolution image from multiple low resolution images when there is no relative motion between the camera and the scene. If there is a small relative motion, the first step to obtain a super-resolved image would be to register the images. This is done by computing the motion of pixels between

image pairs. After the motion is computed, a high resolution image can be obtained by fusing several successive frames covering the same scene. Motion can be calculated using a simple parametric form [29] or by using an optical flow field [30].

Numerous techniques have been proposed for reconstructing a single high-resolution image from multiple low-resolution images of the same scene, which differ by the methods used and assumptions made. The earliest work was by Tsai and Huang [17] in the Fourier domain. A registration approach was proposed which was based on minimizing the energy of the high-resolution signal. In this work, the factors of noise and blur were not considered. Kim et. al [31] extended this work to include noise and blur by using Tikhonov regularization.

In [32], the restoration and interpolation steps were combined together, assuming that the shifts between pixels were known exactly. In [33], the method of projection onto convex sets (POCS) [34] was used to account for noise and blur. Irani and Peleg [35] proposed an iterative technique to estimate the displacements, similar to the back-projection method commonly used in computed tomography. The application of maximum a posteriori (MAP) estimation and POCS for this problem was described in [28]. The technique of applying simultaneous restoration, registration and interpolation was described in [36] by using a MAP framework. Baker and Kanade [37] propose an algorithm for simultaneous estimation of super-resolution video and optical flow taking as input a conventional video stream. A discussion on how this technique is particularly useful for super-resolution of video sequences of faces is provided in the following sections.

2.4 Techniques used in Super-Resolution

As discussed in the above sections, for super-resolution using multiple images, it is necessary to calculate the shifts of pixel intensities occurring in low-resolution images. In most practical applications, in addition to the fact that these shifts are unknown, the low-resolution images might be degraded by blur and noise. Thus, before interpolation to reconstruct the high-resolution image it is necessary to restore the low-resolution images. Therefore, reconstruction of high-resolution images involves the following three tasks: restoration of the low-resolution images, registration of the low resolution images to find the shifts, and interpolation to reconstruct the high-resolution image. These tasks can be

implemented sequentially or simultaneously according to the reconstruction techniques adopted. A brief explanation of a few image processing tasks, used in the process of super-resolution, is provided in the following sections.

2.4.1 Registration

Registration of images is the process of overlaying two or more images by calculating a transformation, which aligns the images together. Given two images I_1 and I_2 where $I_1(x, y)$ and $I_2(x, y)$ represent the intensity values of each image at the location (x, y), the registration of images can be expressed as:

$$I_2(x, y) = g(I_1(f(x, y))) \tag{2.2}$$

where f is a two dimensional spatial coordinate transformation which maps the two spatial coordinates x and y, to new spatial coordinates $x^{'}$ and $y^{'}$, such that

$$(x^{'}, y^{'}) = f(x, y) \tag{2.3}$$

and g is a one dimensional intensity transformation.

Registration is useful in various image processing tasks such as motion analysis, change detection and image fusion. Registration is needed to fuse information contained in two images, which differ due to variation in factors such as time, sensors, viewpoint and position of the object under consideration or motion.

The steps involved in registering two images can be listed as follows:

- Detection of features: First, the most significant features or landmark points are selected either manually or automatically. The selected points could be any of the area based, line based, region based or edge based features present in the image. Shi and Tomasi [38] provide a detailed explanation of selecting good features for registration.

- Matching of features: The selected features can be matched by using either the image intensity values in the close neighborhoods, or the spatial distribution of the features. To speed up the feature matching process, pyramidal approaches can be used [39].

- Estimation of transformation model: After the features are matched, a mapping function which describes the correspondence between the matched features is calculated. The mapping functions can range from simple affine or geometric transformations to complex spline based models [40]. Based on the amount of information used to generate a model, transformation models can be classified into either global or local mapping models. Global models use all the feature points in an image to generate a transformation model where as the local models use feature points in the tessellated image.

- Transformation of images: By using the transformation model or the mapping function calculated by the previous step, images are transformed and registered with each other. Techniques like warping or blending of images are used to fuse the images using the mapping functions.

Figure 2.3 shows an example of image registration. Different techniques used for image registration are listed in [41] and [42]. In the current work, image registration is performed by estimating the difference in intensities between two images by calculating the optical flow.

Figure 2.3: Registration of two images by manual feature selection.

2.4.2 Interpolation

Interpolation refers to the estimation of pixel intensities when an image is resized from a lower resolution to a higher resolution. Interpolation techniques can be mainly classified into two groups: *adaptive* and $non - adaptive$ techniques [43]. Adaptive techniques depend on the intrinsic features of an image like hue, edge information, etc [44]. On the other hand, non-adaptive techniques do not use any intrinsic feature of an image and apply a specific computational logic on the image pixel intensities to interpolate an image. Non-adaptive techniques are attractive and are widely used due to their ease of computation. Some of the non-adaptive techniques are: nearest neighbor, bilinear and bicubic interpolation. Figure 2.4 illustrates an image interpolated with various non-adaptive techniques. A detailed discussion of image interpolation techniques and an evaluation between the performance of each technique is listed in [45]. A survey of interpolation techniques in the medical image processing domain is presented in [46].

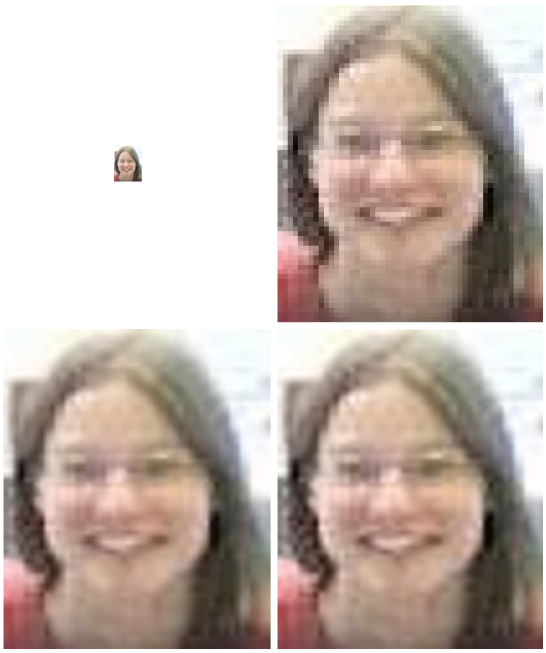

Figure 2.4: Non-adaptive interpolation methods: original image (top left) enlarged ten times using nearest neighbor (top right), bilinear (bottom left), and bicubic (bottom right) interpolation methods, respectively.

2.4.3 Restoration

Image degradation occurs mainly during image acquisition or transmission [47]. Major types of degradation refer to blur, noise, geometrical deformation, etc. Image restoration refers to the process of improving the information content present in a degraded image, by estimating a mathematical model for the degradation.

The goal of restoration is to remove identified distortion from an observed image g, providing a best possible estimate $\overline{f}$ of the original undistorted image f. Major tasks in image restoration are: noise removal and image de-blurring. The degradation, or blurring, of an image can be caused by factors like motion during image capture or out-of-focus optics. A blurred or degraded image can be approximately modeled by the following equation:

$$g = Hf + n, \tag{2.4}$$

where g refers to the blurred image, f refers to the original image, H is the distortion operator, also called the point spread function (PSF) and n is additive noise, introduced during image acquisition, that corrupts the image. In the spatial domain, the PSF describes the degree to which an optical system blurs or spreads a point of light. The importance of the PSF is that based on the modeling of PSF, the task of deblurring can be easily performed. Deblurring of an image can be done by de-convolving the blurred image with the PSF that exactly describes the distortion. The quality of the deblurred image is mainly determined by knowledge of the PSF [48]. A good review of image restoration techniques is provided in [49].

2.5 Applications of Super-Resolution

Super-resolution is a very important process for generating high resolution images in applications such as:

(a) Video surveillance

(b) Video enhancement and restoration

(c) Video standards conversion

(d) Digital mosaicing

(e) Forensic, scientific, medical and satellite imaging.

2.6 Human Faces in Surveillance Videos

Most super-resolution techniques try to register images by using simple parametric transformations such as translations and warps. An assumption is made in most techniques that the objects in the video sequences are rigid. Though the assumption works well in static scenes, the performances of these techniques degrade in the case of human faces in surveillance videos, as human faces are non-planar, non-rigid, non-lambertian and subject to self occlusion [37], [16]. To handle the problem of non-rigidity of faces, authors in [37] suggest allowing the image registration to be an arbitrary flow field. Optical flow techniques can effectively be used to achieve this purpose. The following section discusses optical flow and some of the significant techniques used in its computation.

2.7 Optical Flow

The process of estimating motion in time-ordered image sequences as either instantaneous image velocities or discrete image displacements can be referred as optical flow field estimation [50]. Optical flow estimation is very useful in performing motion detection, image mosaicing, image reconstruction and object segmentation tasks. Following are the major classes of optical flow estimation [51]:

(a) Differential techniques which compute velocity from spatiotemporal derivatives of image intensity or filtered versions of the image using low or band pass filters.

(b) Region based techniques which consider maximizing a similarity measure or minimizing a distance measure and compute the velocities using region based matching.

(c) Energy based techniques which are based on the the function of output energy of velocity tuned filters in the fourier domain.

(d) Phase based techniques which estimate velocity in terms of phase behavior of band pass filter outputs.

The following major assumptions are taken into account, when optical flow between two images is to be to be computed:

(a) Brightness is constant across the set of frames in consideration.

(b) Motion occuring in two consecutive frames is small.

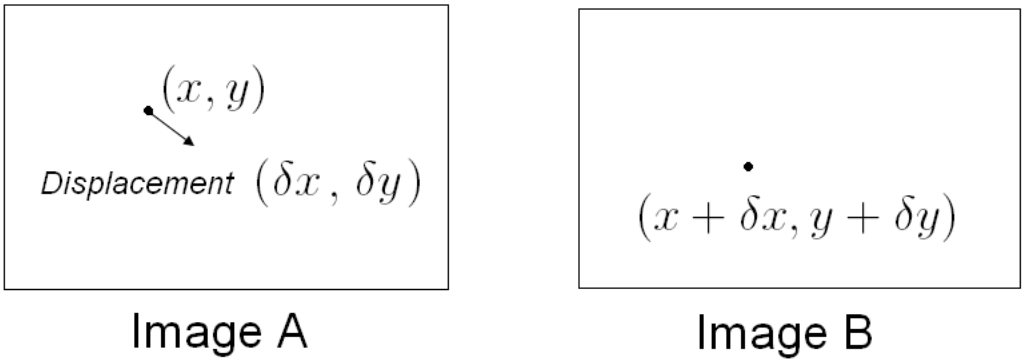

Figure 2.5: Displacement of a pixel intensity in two images.

Consider two images A and B, in which the pixel intensity located at (x, y), is displaced by δx and δy along the horizontal and the vertical axes. By considering the brightness constancy assumption, the pixel intensities in the two images can be related as follows:

$$I^A(x, y) = I^B(x + \delta x, y + \delta y). \tag{2.5}$$

By using the Taylor series expansion, intensity of the pixel in the second image can be expressed as following:

$$I^B(x + \delta x, y + \delta y) = I^B(x, y) + \frac{\partial I^B}{\partial x}\delta x + \frac{\partial I^B}{\partial y}\delta y + \text{higher order terms.} \tag{2.6}$$

By neglecting the higher order terms, the equation can be written as:

$$I^B(x + \delta x, y + \delta y) \approx I^B(x, y) + \frac{\partial I^B}{\partial x}\delta x + \frac{\partial I^B}{\partial y}\delta y. \tag{2.7}$$

From equation (2.5), it can be written as:

$$I^B(x + \delta x, y + \delta y) - I^A(x, y) = 0. \tag{2.8}$$

By using equation (2.7), we get:

$$I^B(x,y) + I_x\delta x + I_y\delta y - I^A(x,y) \approx 0, \tag{2.9}$$

where $I_x = \frac{\partial I^B}{\partial x}$ and $I_y = \frac{\partial I^B}{\partial y}$. Now, the above equation can be modified as follows:

$$(I^B(x,y) - I^A(x,y)) + I_x\delta x + I_y\delta y \approx 0, \tag{2.10}$$

$$I_t + I_x\delta x + I_y\delta y \approx 0. \tag{2.11}$$

Thus, the *classical optic flow constraint equation* can be written as:

$$I_x u + I_y v + I_t \approx 0, \tag{2.12}$$

where I represents the intensity of the pixel under consideration and (u, v) respectively represent the horizontal and vertical components of the flow, and the subscripts stand for differentiation. The equation is not sufficient to determine (u, v) and, thus, optic flow computation by this equation is an under constrained problem.

The optic flow equation can be solved by imposing additional constraints. The conditions applied to solve the equation vary with the optical technique considered. Optical flow techniques can be classified into *global* or *local* techniques, based on the manner in which the conditions are imposed.

Global techniques usually impose constraints over the whole image space whereas local techniques exploit information from a small neighborhood around an examined pixel. The imposition of constraints over the whole image space, in global techniques, introduces a correlation among the field vectors that might not have any physical reason across motion boundaries. Thus, global techniques usually produce incorrect flattening of the computed field. Though improvements have been suggested for global techniques, they are computationally demanding [52].

On the other hand, local techniques are simple and fast. As these techniques do not impose any smoothness over large patches of images, they offer the advantage of eliminating any undesirable flattening effects of the motion field.

One very popular local technique suggested by Lucas and Kanade [53] is considered for optical flow calculations and registration of images in this work.

2.8 Lucas-Kanade Technique

According to the discussion in the previous section, Lucas-Kanade method divides the original image into smaller sections and assumes a constant velocity in each section. Thus, the optical flow equation for a pixel p_i, which is the center of a window, can be written as:

$$I_t(p_i) + \nabla I(p_i).\begin{bmatrix} u \\ v \end{bmatrix} = 0. \tag{2.13}$$

If a window of size 3×3 is assumed, a set of 9 equations can be obtained per pixel, which can be written in matrix form as:

$$\begin{bmatrix} I_x(p_1) & I_y(p_1) \\ I_x(p_2) & I_y(p_2) \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ I_x(p_9) & I_y(p_9) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_t(p_1) \\ I_t(p_2) \\ \cdot \\ \cdot \\ \cdot \\ I_t(p_9) \end{bmatrix}. \tag{2.14}$$

The above equation can be written in a simplified manner as follows:

$$Ad = b, \tag{2.15}$$

where $A = \begin{bmatrix} I_x(p_1) & I_y(p_1) \\ I_x(p_2) & I_y(p_2) \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ I_x(p_9) & I_y(p_9) \end{bmatrix}$, $d = \begin{bmatrix} u \\ v \end{bmatrix}$ and $b = - \begin{bmatrix} I_t(p_1) \\ I_t(p_2) \\ \cdot \\ \cdot \\ \cdot \\ I_t(p_9) \end{bmatrix}$. Solution for equation (2.15) can be achieved by solving the least squares problem, which leads to the following equation:

$$(A^T A)d = A^T b. \tag{2.16}$$

Thus, we get:

$$\begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix}. \tag{2.17}$$

The above equation can be solved for the values of u and v, the displacements of a pixel along the vertical and horizontal axes, by considering:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum I_x I_t \\ -\sum I_y I_t \end{bmatrix}. \tag{2.18}$$

Chapter 3

Super-Resolution Optical Flow

3.1 Introduction

Super-resolution optical flow, first described in [37], is a technique for simultaneously computing both the optical flow and a higher resolution version of an entire video. This spatial domain algorithm is very useful in processing human surveillance videos due to its:

- ability to accurately register motion between successive frames in a given video, and
- applicability to face videos, as it allows image registration to be an arbitrary flow field.

Like all other super-resolution algorithms, super-resolution optical flow technique has four major steps towards obtaining a high resolution image: Registration, Warping, Fusion and Deblurring [37]. The strength of super-resolution optical flow lies in the fact that it combines the registration and fusion steps, estimating both registration and super-resolution of the images at the same time.

3.2 Super-Resolution Optical Flow Algorithm

Let $\overline{V}$ be a low resolution video sequence and $F = \{f_1, f_2, f_3, ..., f_n\}$, denote the n frames constituting $\overline{V}$. To obtain a super-resolution version of the i^{th} frame, $f_i^{'}$, super-resolution optical flow algorithm utilizes a set of frames $f_{i-2}, f_{i-1}, f_i, f_{i+1}$ and f_{i+2}. The steps involved in the execution of the algorithm are as follows:

(1) Frames $\{b_1, b_2, b_3, ..., b_n\}$, having twice the resolution of the original frames are obtained by bilinearly interpolating $\{f_1, f_2, f_3, ..., f_n\}$, respectively.

(2) The optical flow fields relating the frame b_i with frames b_{i-2}, b_{i-1}, b_{i+1} and b_{i+2} are computed.

(3) Using the calculated optical flow, b_{i-2} and b_{i-1} are warped 'forward' while b_{i+1} and b_{i+2} are warped 'backward' into the coordinate frame of b_i to obtain the warped frames $w_{i-2,i}$, $w_{i-1,i}$, $w_{i,i+1}$ and $w_{i,i+2}$ respectively.

(4) The four warped frames are blended with b_i using robust mean calculations and the resulting image is deblurred by a Wiener deconvolution filter to obtain a final super-resolution image, f_i'.

This process is repeated for all the frames in the video sequence, starting from f_3 till f_{n-3}. Figure 3.1 illustrates the technique using a flow diagram.

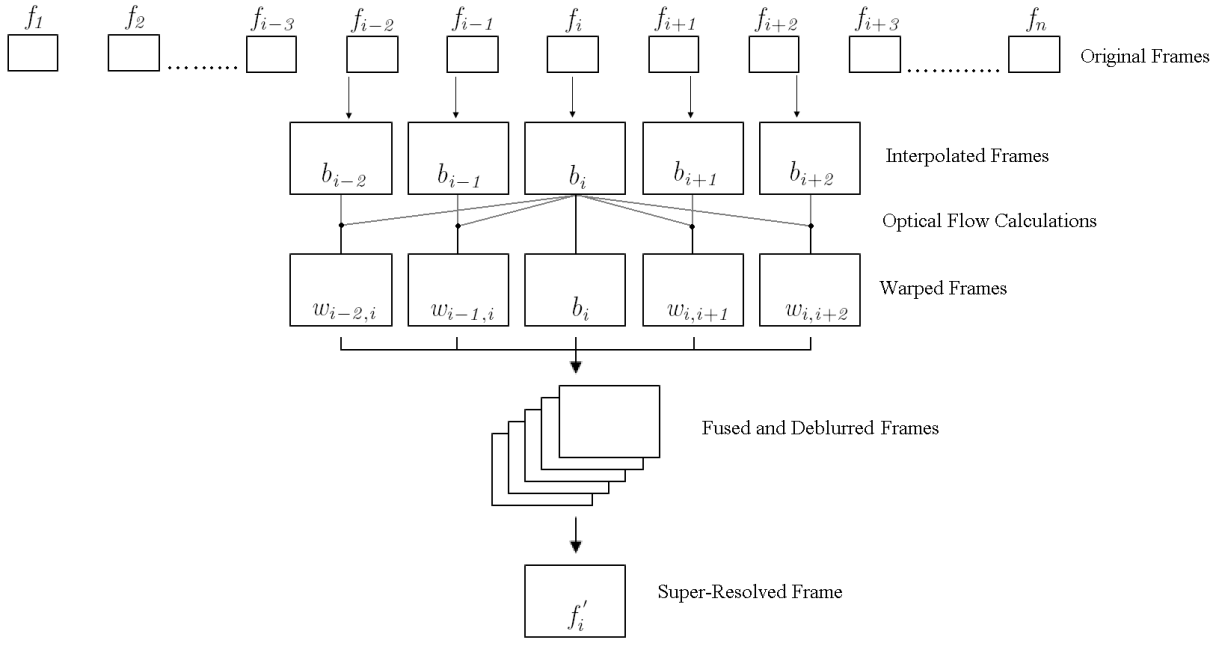

Figure 3.1: Flow diagram of super resolution optic flow.

The technique was originally discussed in [54]. This algorithm has been evaluated on various databases in [55] and [56].

3.3 Generalized Version

The original version of super-resolution optical flow [37] considers five successive frames to generate a super resolved image. Given a video with n frames, starting from third frame and continuing until the third frame from last, this technique can effectively generate $(n-3)$ super-resolved frames.

In [56], the same algorithm was implemented by considering (2k + 1) frames for generating a high resolution frame, where the value of k was varied as k = 1, 2, 3 and 4. A brief description of the algorithm considering k = 1 is provided here. The process for obtaining a super-resolved image in this case remains the same as the original super-resolution optical flow algorithm, except for the number of frames used. For a given frame f_i, its super-resolved version $f_i^{'}$ is obtained by considering the frames f_{i-1}, f_i and f_{i+1}. The algorithm is illustrated in Figure 3.2.

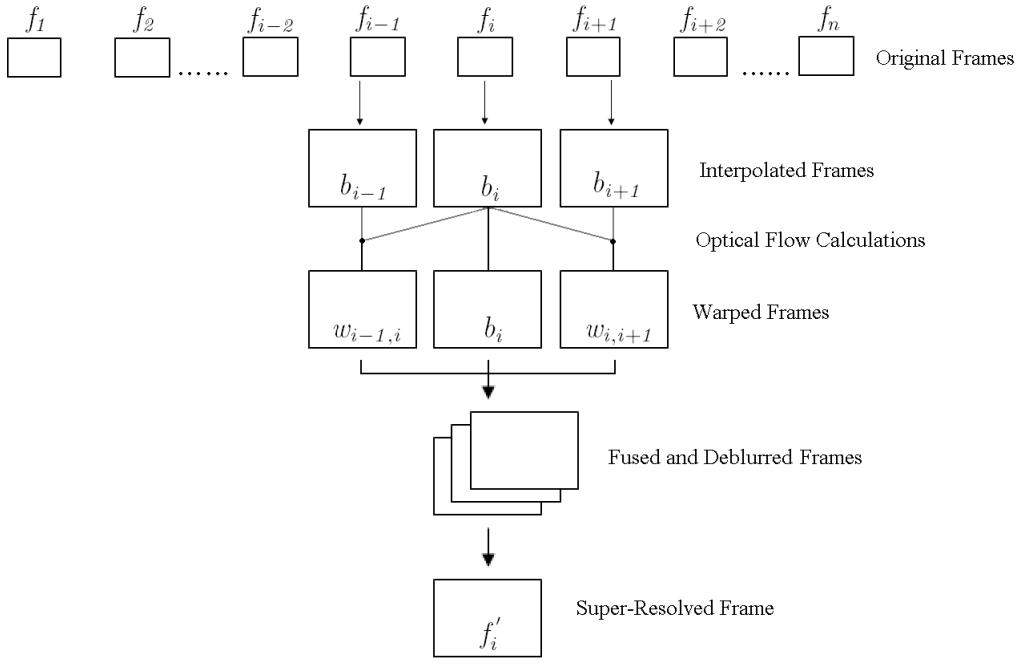

Figure 3.2: Super-resolution optic flow algorithm with k=1.

3.4 Impact of the number of frames

Generally, the performance evaluation of a super-resolution technique is based on the quality of the super-resolved frame and the number of reconstruction errors present in it. Quality assessment refers to objective quality measurement by using quality metrics and visual evaluation. A good super-resolution technique often outputs a higher quality frame with the least number of reconstruction or estimation errors. In [56], a discussion on the effect of variation of number of frames used for reconstruction on the output is presented. The authors suggest that by considering *k=2* (i.e., 5 frames), an optimal value for trading reconstruction quality with computation time, a super-resolved frame with minimal reconstruction error can be obtained.

Implementation and performance of super-resolution optical flow technique with *k=1* and *k=2* (using 3 and 5 frames, respectively) is studied in this thesis. Quality assessment for evaluating the performance of super-resolution optical flow using *k=2* and *k=1* was done visually.

Visual evaluation is performed by a human observer, manually inspecting a super-resolved frame and rating the performance of the technique used to produce the frame. Since visual evaluation is a time consuming process, no statistics were recorded using human observers in this thesis. A short inspection carried out by a human observer reveals that not all frames in the sets generated by using *k=1* and *k=2* are of high quality.

Certain frames are corrupted by noise and contain artifacts, which alter the appearance of an individual in the frame. A difference in the outputs by varying the image reconstruction technique can be seen in Figure 3.3.

Figure 3.3: A closer look at the super-resolution frames reconstructed using k=2 and k=1.

3.5 Artifacts or Reconstruction Errors

Artifacts are a group of noisy pixels induced in a high resolution frame due to incorrect registration between two input frames. They can range from minor pixel value estimation errors to completely degraded frames which can heavily alter the information content in a frame. Artifacts in a frame occur due to incorrect registration which is caused by a large displacement due to motion in a scene, or by noise in the system.

If large motion occurs in a short frame of time, aligning the corresponding frames might cause reconstruction errors resulting in artifacts. A detailed discussion about the extent of reconstruction errors in high-resolution frames generated by using super-resolution optical flow with *k=1* and *k=2* and their causes are presented below.

Visual evaluation of the reconstructed frames reveals that the number of artifacts occurring in frames generated with $k = 2$ is more compared to that of k=1. It was also observed that the number of frames affected by artifacts were high in the case of k=2. The effect of image degradation was due to the impact of the frames succeeding and preceding the frame which exhibited large motion.

Suppose that in a set of low resolution frames $\{f_k,\ f_{k+1},\ f_{k+2}, f_{k+3}, f_{k+4}, f_{k+5}\}$, the frames f_{k+2} and f_{k+3} have a large inter-frame motion, then not only do the reconstructed frames f'_{k+2} and f'_{k+3} suffered degradation, but also the frames f'_k till f'_{k+5} show estimation errors. The reason for such effect is the repeated usage of a frame, with large pixel displacements, during the reconstruction process. Since a given frame is considered a large number of times during the reconstruction process for k=2, the artifacts are seen in more number of output frames. Figures 3.4 and 3.5 illustrate the discussed effect.

Figure 3.4: Low resolution frames used for super-resolution process.

Since presence of artifacts hinders the recognition performance; it is desirable to eliminate artifacts in the output frames. To achieve artifact elimination, a technique is needed which first discards the frames with large motion. This can be achieved by adaptively selecting frames, which forms a basis for the Adaptive Frame Selection (AFS) technique.

Figure 3.5: Artifacts in frames reconstructed using k=2 and k=1.

3.6 Adaptive Frame Selection Technique

The adaptive frame selection technique aims to overcome the registration errors caused by inter-frame motion and to improve the performance of the super-resolution optical flow technique. The purpose of this technique is to adaptively choose the frames in a given video sequence for the reconstruction process based on the motion occurring between two consecutive frames. The main features of this algorithm are:

- quantification of motion between frames, and
- frame selection, which would be used for super-resolution frame reconstruction.

3.7 Inter-Frame Motion Parameter

A technique to quantify the motion between two consecutive frames by utilizing the information from optical flow calculations is presented in this section. A parameter called Inter-Frame Motion Parameter, β, which records the motion between two consecutive frames is proposed. Assume two consecutive frames, f_k and f_{k+1}, in a given video sequence $\overline{V}$, both having a resolution of $(M \times N)$ pixels. Let the intensities of pixels in f_k and

f_{k+1} be represented by the equations (3.1) and (3.2), respectively.

$$I_{f_k} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdot & \cdot & \cdot & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdot & \cdot & \cdot & a_{2,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{M,1} & a_{M,2} & \cdot & \cdot & \cdot & a_{M,N} \end{bmatrix}, \tag{3.1}$$

$$I_{f_{k+1}} = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdot & \cdot & \cdot & b_{1,N} \\ b_{2,1} & b_{2,2} & \cdot & \cdot & \cdot & b_{2,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ b_{M,1} & b_{M,2} & \cdot & \cdot & \cdot & b_{M,N} \end{bmatrix}. \tag{3.2}$$

By using the Lucas-Kanade algorithm, the optical flow matrices are calculated. The optical flow matrices $\underline{u}_{k,k+1}$ and $\underline{v}_{k,k+1}$, which represent the individual pixel displacements of the pixels in the frames f_k and f_{k+1} in both x and y directions , can be represented as:

$$\underline{u}_{k,k+1} = \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdot & \cdot & \cdot & \delta_{1,N} \\ \delta_{2,1} & \delta_{2,1} & \cdot & \cdot & \cdot & \delta_{2,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \delta_{M,1} & \delta_{M,2} & \cdot & \cdot & \cdot & \delta_{M,N} \end{bmatrix}, \tag{3.3}$$

$$\underline{v}_{k,k+1} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdot & \cdot & \cdot & \sigma_{1,N} \\ \sigma_{2,1} & \sigma_{2,1} & \cdot & \cdot & \cdot & \sigma_{2,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{M,1} & \sigma_{M,2} & \cdot & \cdot & \cdot & \sigma_{M,N} \end{bmatrix}, \tag{3.4}$$

where δ_{pq} and σ_{pq} denote the displacements between the frames I_{f_k} and $I_{f_{k+1}}$ at the pixel location (p, q) along x and y directions, respectively.

After calculating the optical flow, the flow magnitude matrix $\overline{\mathbb{L}}$ is populated by con-

sidering the L2 norm of the displacements along both axes at each pixel. Thus we have,

$$\overline{\mathbb{L}} = \{\overline{L}_{p,q}\}, \tag{3.5}$$

for $p \in P$ *and* $q \in Q$ *and* p=*{1,2,3,...,M}*, q=*{1,2,3,...,N}*, where

$$\overline{L}_{p,q} = \|\delta_{p,q} - \sigma_{p,q}\|_2. \tag{3.6}$$

Once the flow magnitude matrix $\overline{\mathbb{L}}$ is calculated, the mean of the top m values of the matrix, sorted in descending order, represents β. For every consecutive pair of frames, f_k and f_{k+1} in the sequence, $\beta(f_k, f_{k+1})$ is computed. All β values are then normalized by using the min-max rule to transform the data to a new range, generally [0,1].

3.8 Threshold Value

After obtaining the inter-frame motion values for the entire video, a threshold T is decided which is used for adaptive selection of frames. Selecting the value for T is an important task as it helps detect the frames possessing large inter-frame motion values. The value of T in our experiments is decided empirically.

All successive frames whose *beta* values fall below T would be used for reconstruction process. Since the frames with β values greater than T are considered to have high motion, they could be thought of as frames which contain multiple poses of the same subject. This information could later be used for selecting frames pertaining to an individual's face recorded at different pose angles. Figures 3.6 and 3.7 describe the process and the algorithm of adaptive fusion technique, respectively.

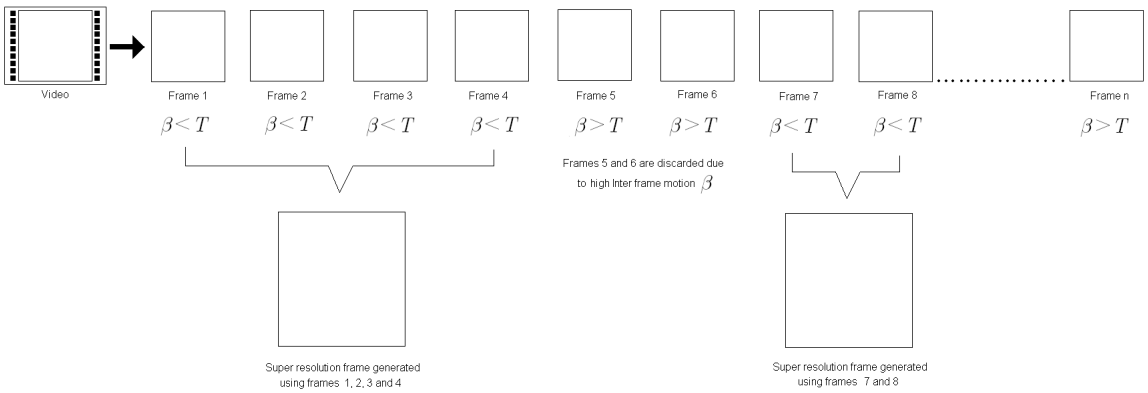

Figure 3.6: Flow chart for the proposed adaptive fusion technique.

Input:
n low resolution frames

Variables:
n is the number of frames in a given video
β is the inter frame motion parameter for the given frame
T is the threshold set to drop frames
k, count are counters

Algorithm:

```
k = 1;
while (k < = n)
        if β(f_k) < T
                frame f_k considered for fusion;
                count = count + 1;
        else
                frame f_k dropped;
                if count = 1
                        bilinearly interpolate and save the frame;
                else
                        if count < 5
                                fuse all the considered frames;
                        elseif count > 5
                                fuse frames in sets of 5;
                        end
                        count = 0;
                end
        end
        k=k+1;
end
```

Output:
super-resolved frames

Figure 3.7: Algorithm for adaptive fusion technique.

Chapter 4

Image Quality

4.1 Introduction

Image quality assessment refers to determining the degree or grade of merit which can be assigned to an image with respect to the amount of useful information present in it. Image quality assessment plays a key role in evaluating and optimizing image processing systems [57]. Given an image quality assessment system, it is expected that the measures used to assess the quality would follow visual perception characteristics.

Image quality metrics are of a significant importance in image processing applications as they can be used to:

(a) monitor and adjust the quality of an image under consideration. For example, during data collection in biometrics, images can be examined for quality before storing them in the database.

(b) optimize image processing algorithms. For example, algorithms can be designed based on image quality, which consider only good quality images for processing.

(c) compare and benchmark various image processing algorithms. For example, the performances of various algorithms can be compared, by evaluating the quality of the image reconstructed by them.

4.2 Image Quality Metrics

Based on the technique used, quality assessment can be classified into two groups: Subjective and Quantitative [58].

4.2.1 Subjective criteria or Human Visual System characteristics

If the final user of images are humans, most reliable assessment of image quality can be achieved by subjective rating of images by human observers. This task can be achieved in two ways: *absolute evaluation*, where an individual assesses the quality of an image by assigning to it a category in a given rating scale, or *comparative evaluation*, where a set of images are ranked from best to worst. Once an image is rated, the mean rating of a group of observers who evaluate the images is usually computed by the following equation:

$$R = \frac{\sum_{k=1}^{n} s_k n_k}{\sum_{k=1}^{n} n_k}, \tag{4.1}$$

where s_k is the score corresponding to the k^{th} rating and n_k refers to the number of observers with that rating and n represents the number of grades on the scale.

4.2.2 Quantitative criteria or Mathematically defined measures

As the name suggests, mathematically defined measures are employed for the task of image quality assessment in this class. Quantitative measures for image quality can be divided into two classes - *univariate* and *bivariate* - which are explained in the following sections. Graphical measures like histograms and Hosaka plots [59] can also be used for image quality measurement.

4.2.3 Significance of Quantitative Criteria

Subjective rating is affected by a number of criteria like viewing angles, level of expertise of the observers, and the type and range of images. With different criteria, subjective

evaluation might produce different measurements results that are inconvenient to use. Also, it is observed that none of these complicated metrics have shown any clear advantage over the quantitative criterion [60].

Quantitative measures, on the other hand, are attractive as they are inexpensive, easy to calculate with low computational complexity, and are independent of viewing conditions of individual observers. Also, as they are standardized calculations, results obtained in different locations at different times are comparable.

4.3 Univariate Measures

Univariate measures assign a numerical value to a single image based upon the measurements of an image field. If the original image field in spatial domain can be represented as $F(j,k)$ with a resolution of $M \times N$ pixels, then a univariate quality rating may be expressed in general by equation (4.2):

$$Q = \sum_{j=1}^{M} \sum_{k=1}^{N} O\{F(j,k)\}, \tag{4.2}$$

where O is any operator considered to compute the desired measure.

If z_i represents a random variable indicating intensity in an image, $p(z)$ the histogram of the intensity levels in a region, L the number of possible intensity levels, then a sample list of univariate measures which calculate image quality based on textural content [61] are discussed below:

- Mean: Mean is the measure of average intensity in an image, represented by equation (4.3):

$$m = \sum_{i=0}^{L-1} z_i p(z_i). \tag{4.3}$$

- n^{th} moment: The n^{th} moment about the mean can be represented by equation (4.4):

$$\mu_n = \sum_{i=0}^{L-1} (z_i - m)^n p(z_i). \tag{4.4}$$

The n^{th} moment represents the variance of the intensity distribution for $n = 2$ and

skewness of the histogram for $n = 3$.

- Standard deviation: A measure of average contrast in an image is represented by standard deviation given by equation (4.5):

$$\sigma = \sqrt{\mu_2(z)}. \tag{4.5}$$

- Smoothness: The relative smoothness of the intensity in a region can be measured by equation (4.6):

$$R = 1 - (1/(1 + \sigma^2)). \tag{4.6}$$

  The value of R is 0 for a region of constant intensity and approaches 1 for regions with large excursions in the values of its intensity levels.

- Uniformity: A measure of uniformity of the intensity values in an image can be given by equation (4.7):

$$U = \sum_{i=0}^{L-1} p^2(z_i). \tag{4.7}$$

- Entropy: The randomness of the intensity values in a given image can be measured by calculating the entropy, which is represented by equation (4.8):

$$e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i). \tag{4.8}$$

4.4 Bivariate Measures

Bivariate measures output a measure of quality by numerically comparing two different images, with one as a reference. These measures can be defined in either spatial or frequency domain. Based on the availability of a reference image, bivariate quality measurement approaches can be divided into *full reference techniques*, where a complete reference image is assumed to be known, and *no reference* or *blind quality assessment techniques*, where a reference image is not available.

If $F(j, k)$ and $F^{'}(j, k)$ denote the original and reconstructed images with a resolution of $M \times N$, a number of measures can be established to determine the closeness of the two images:

- Average Difference:

$$AD = \sum_{j=1}^{M}\sum_{k=1}^{N}[F(j,k) - F^{'}(j,k)]/MN. \tag{4.9}$$

- Structural Content:

$$SC = \frac{\sum_{j=1}^{M}\sum_{k=1}^{N}[F(j,k)]^2}{\sum_{j=1}^{M}\sum_{k=1}^{N}[F^{'}(j,k)]^2}. \tag{4.10}$$

- Normalized Cross Correlation:

$$NK = \frac{\sum_{j=1}^{M}\sum_{k=1}^{N}F(j,k)F^{'}(j,k)}{\sum_{j=1}^{M}\sum_{k=1}^{N}[F(j,k)]^2}. \tag{4.11}$$

- Correlation Quality:

$$CQ = \frac{\sum_{j=1}^{M}\sum_{k=1}^{N}F(j,k)F^{'}(j,k)}{\sum_{j=1}^{M}\sum_{k=1}^{N}F(j,k)}. \tag{4.12}$$

- Maximum Difference:

$$MD = Max\{|F(j,k) - F^{'}(j,k)|\}. \tag{4.13}$$

- Image Fidelity:

$$IF = 1 - (\frac{\sum_{j=1}^{M}\sum_{k=1}^{N}[F(j,k) - F^{'}(j,k)]^2}{\sum_{j=1}^{M}\sum_{k=1}^{N}[F(j,k)]^2}). \tag{4.14}$$

- Peak Mean Square Error:

$$PMSE = \frac{1}{MN} \frac{\sum_{j=1}^{M} \sum_{k=1}^{N} [F(j,k) - F^{'}(j,k)]^2}{[Max\{F(j,k)\}]^2}. \tag{4.15}$$

- Normalized Absolute Error:

$$NAE = \frac{\sum_{j=1}^{M} \sum_{k=1}^{N} F(j,k) - F^{'}(j,k)}{|F(j,k)|}. \tag{4.16}$$

- Normalized Mean Square Error:

$$NMSE = \frac{\sum_{j=1}^{M} \sum_{k=1}^{N} [F(j,k) - F^{'}(j,k)]^2}{\sum_{j=1}^{M} \sum_{k=1}^{N} [F(j,k)]^2}. \tag{4.17}$$

- L_p Norm:

$$L_p = \{\frac{1}{MN} \sum_{j=1}^{M} \sum_{k=1}^{N} |F(j,k) - F^{'}(j,k)|^p\}^{1/p} \ , \ p = 1, 2, 3. \tag{4.18}$$

An extensive survey and classification of quality measures was provided by Eskicioglu and Fisher [58]. A performance evaluation of quality measures was provided in [60] with an indication of NMSE, IF, NAE and Lp being the best set of measures to use for quality assessment. An extended list of image quality measures was presented in [62] by classifying measures based on various criteria. A universal quality index was proposed by Wang and Bovik [63], which does not depend on the images being tested or the viewing conditions. The system was trained using a set of good and bad images which was later used to grade previously unseen images based. No reference quality assessment was described in [64] and [65]. An image quality assessment based on structural similarity of images was proposed by Wang et. al [66]. A system to exclusively assess the quality of facial images was described in [67] but the metrics used were dependent more on the scene

and photographic conditions rather than intrinsic image properties. Although a lot of research has been done to propose new quality metrics, it is always noted that the quality assessment largely depends on the application.

In this thesis, the image quality metrics were considered initially for the purpose of frame selection for fusion. It was observed that the univariate measures which captured the textural content of an image, were not effective in expressing the image degradation due to motion. Also, it was noticed that the bivariate image quality metrics did not reflect variations in magnitude which would be helpful in comparing images based on motion degradation. Thus, the image quality metrics were only considered whilst comparing the performances of the super-resolution techniques. For this purpose, Mean Square Error (MSE) was considered. MSE is the simplest and most widely used quality metric. Though its correlation with visual quality is low, MSE is appealing due to the following reasons:

- Simplicity of calculation.
- Clear physical meaning.
- Mathematically convenient in the context of optimization.

Chapter 5

Results

5.1 Database

The super-resolution techniques described in this work can be applied to any video sequence, but the research is mainly oriented towards extracting facial information from surveillance videos. Thus, a database which contains videos comparable to surveillance videos is required for performance evaluation purposes. To meet the requirement, the IIT-NRC facial video database [68] is considered in this thesis. Factors such as *low resolution, motion blur, out-of focus factor, variations in facial expressions and orientations* and *occlusions without being affected by illumination*, make the videos of the database comparable to surveillance videos.

The database contains a set of brief, low-resolution video clips, each showing the face of a user sitting in front of a computer. Users exhibit a wide range of facial expressions and orientations, which are captured by a USB webcam mounted on the computer monitor. The video capture size is maintained at 160×120 pixels resulting in an inter-pupillary distance of about 12 pixels in the video. The database consists of two recordings for each of eleven individuals resulting in 22 videos, which were compressed with the AVI Intel codec.

Since the processing and evaluation of the techniques described in this work are frame-based, all the videos are first converted to sequences of frames. Four videos of the database, which are of a higher resolution compared to the rest of the dataset, have been resized to 160×120 pixels to maintain uniformity in evaluation. With each video clip recorded at a

rate of 20 frames per second and having a time duration of 10 to 15 seconds, the number of frames extracted from all the videos is over 6900.

For evaluating the performance of the techniques considered in this work, two evaluation schemes have been considered: *quality metrics-based evaluation* and *match score-based evaluation*. These evaluation schemes are explained in detail in Sections 5.2 and 5.4, respectively.

5.2 Quality Metrics-based Evaluation

The evaluation based on quality metrics, compares the performance of different techniques using Mean Square Error (MSE), the inter-frame motion parameter β, and the image quality metrics defined in the previous chapter. To compare the quality of a frame using some of the bivariate measures mentioned in Section 4.4, it is necessary to have a reference frame for comparison purposes.

Due to the very low resolution of the frames extracted from the videos and the lack of any ground truth data, it was necessary to resize the low resolution frames to the size of super-resolved frames. This was achieved by resizing all the frames using bi-cubic interpolation, forming a *reference* set. A reference set helps in forming a benchmark dataset, the performance of which can be used to compare the performances of the proposed technique. The process of selecting a reference frame, which corresponds exactly to a given super-resolved frame, is explained in the following section.

5.3 Reference Frames

Consider an SR frame $f_k^{'}$, generated by using the super-resolution optical flow technique on 3 frames (SR3). If $f_k^{'}$ is obtained by combining the information content present in the LR frames $\{f_{k-1}, f_k, f_{k+1}\}$, then the interpolated frame b_k in the reference set is considered as the reference frame for $f_k^{'}$. A similar consideration can be applied to an SR frame generated by the super-resolution optical flow technique using 5 frames (SR5). The interpolated frame b_p is considered the reference frame for the SR frame $f_p^{'}$ that is generated by combining the information content present in the LR frames $\{f_{p-2}, f_{p-1}, f_p, f_{p+1}, f_{p+2}\}$. Figure 5.1 illustrates the discussed concept.

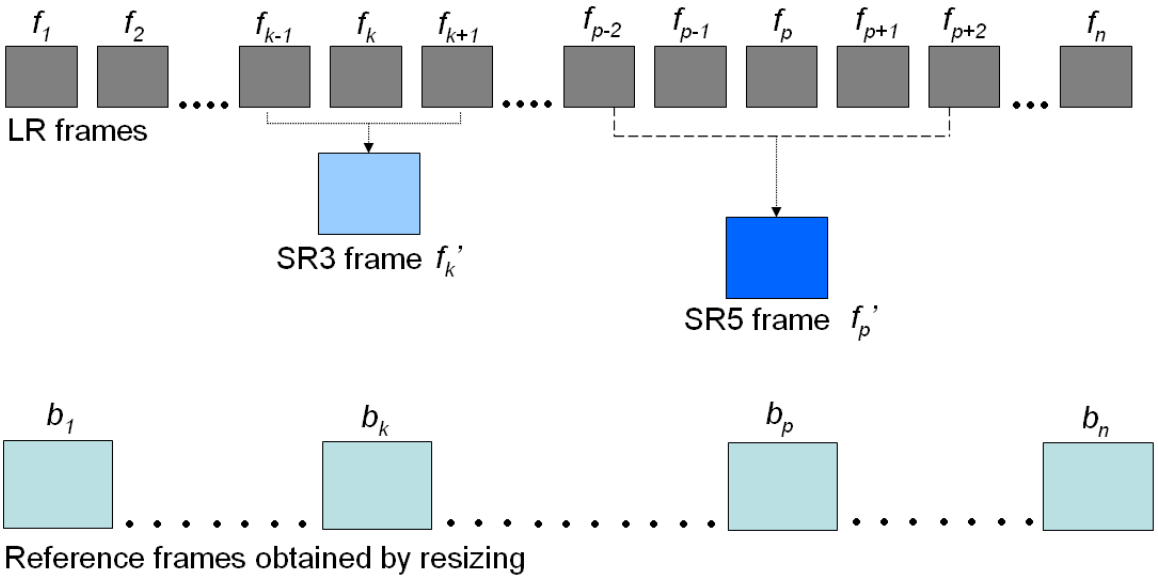

Figure 5.1: Reference frame set for SR3 and SR5 sets.

As discussed earlier, in the case of the adaptive frame selection (AFS) technique it is possible to obtain an SR frame either by interpolation of a single LR frame, or by combining multiple LR frames. If an SR frame is obtained from a single frame f_k, the interpolated frame b_k serves as the reference frame. On the other hand, if a frame is obtained by combining the information in the LR frames $\{f_k, f_{k+1}, ..., f_p\}$, then the interpolated frame b_r is used as reference, where r could be either $\lfloor (k+p)/2 \rfloor$ or $\lceil (k+p)/2 \rceil$.

5.4 Match Score based Evaluation

It was suggested in [69] that the Receiver Operating Characteristic (ROC) curves can be effectively used for the performance evaluation of video surveillance systems. In this thesis, performance evaluation is reported using ROC curves and hit rates. Some of the preliminary tasks required for such evaluation are listed in the following sections.

5.4.1 Gallery and Probe Sets

The frames in the database (gallery) that correspond to the distinctive identity of an individual are considered as gallery frames. Creation of the gallery is performed by considering frames which contain maximum facial information. Such frames are selected

by the following visual criteria: *availability of full-frontal facial pose of the individual, no motion blur* and *no occlusions*.

The frames which need to be compared with gallery frames, to verify the identity of an individual present in them, are considered as probe frames. The verification of identity is based on the match score generated by a comparison of the two frames (probe with gallery). In this work, all the SR frames generated by the various techniques are considered as probe frames.

5.4.2 Template Generation

Template generation is the process of extracting feature sets of interest from an image and associating a label to the identity of an individual, before storing it in a database. These templates could later be used for comparing probe frames and establish their identities. Template generation occurs only if a face can be successfully detected and the features can be extracted from a considered frame. Main reasons for a failure to generate the templates are: motion blur, artifacts, large variation in pose or orientation, occlusions and a large distance from the camera. In this work, the task of template generation was performed using the Identix G6 FaceIt SDK [70].

5.4.3 Score Generation

For this database, given the set of videos $\{V_1, V_2, V_3, V_4, ..., V_{22}\}$, the set of frames extracted from a given video V_k can be represented by $F^k = \{f_1^k, f_2^k, f_3^k, ..., f_p^k\}$, where p denotes the number of frames extracted from the video. It is to be noted that the number of frames extracted from a given video is not constant for all the videos as they are not of the same time duration. The gallery frames which represent the identities present in the database are denoted by $\{G_1, G_2, G_3, ..., G_{22}\}$. For every i^{th} frame in a given video V_j, denoted by f_i^j, a score is generated by comparing it with a given gallery frame G_r. The score S for such comparison is denoted by $S(G_r, f_i^j)$. The Identix G6 FaceIt SDK [70] was used for score generation.

5.4.4 Receiver Operating Characteristic Curves

A score $S(G_r, f_i^j)$ is called a genuine score if it is generated by comparing a probe frame with a gallery frame containing the true identity of the individual present in the frame. All other comparisons yield impostor scores. If a genuine score falls below a specified threshold it is counted as a false reject. On the other hand, if an impostor score exceeds the threshold, it is counted as a false accept. For varying thresholds, False Accept Rate (FAR) and False Reject Rate (FRR) are calculated by the following expressions:

$$FAR = \frac{Number\ of\ false\ accepts}{Number\ of\ impostor\ scores}, \tag{5.1}$$

$$FRR = \frac{Number\ of\ false\ rejects}{Number\ of\ genuine\ scores}, \tag{5.2}$$

When the FAR and FRR error rates are plotted for various threshold values, the resulting curve is called a Receiver Operating Characteristic (ROC) curve.

5.4.5 Identification

In the identification process, the scores obtained by comparing a given frame f_k with all the gallery frames, are first sorted in descending order. Then, if one of the top K scores of the sorted score set relate to the true identity of the individual present in that frame, then it is labeled as a *hit*, else a *miss*. Once the number of hits for all the frames in a video are available, the *hit rate* is calculated by the following equation:

$$Hit\ Rate = \frac{Number\ of\ hits\ in\ a\ given\ video}{Number\ of\ total\ frames\ in\ the\ video}. \tag{5.3}$$

5.5 Fusion

Based on the type and amount of biometric data available, information fusion, which improves performance of a biometric system, can be carried out at multiple levels. Figure 5.2 shows the various levels of fusion possible in a biometric system. A detailed description of various fusion schemes possible in a biometric system and their classifications are provided in [71]. In this work, two significant levels, *image-level* and *score-level* fusion schemes are considered.

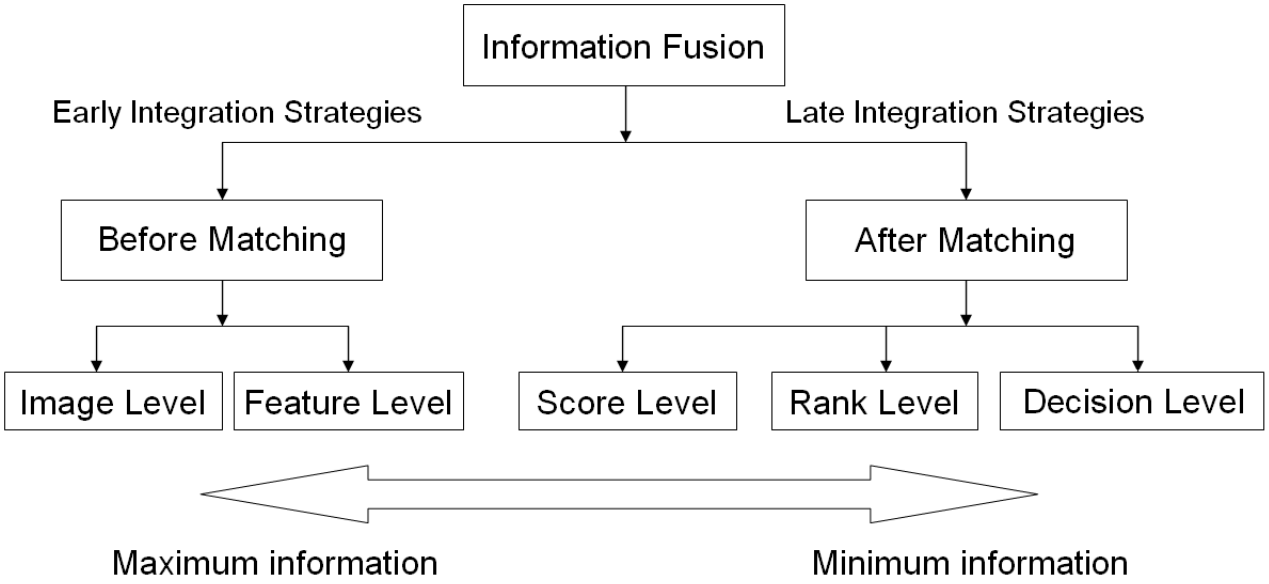

Figure 5.2: Various levels of fusion possible in a biometric system.

5.5.1 Image-Level Fusion

As it is usually considered that the raw biometric data obtained from an individual contains maximum information, it can be expected to improve performance by fusion of the data at the sensor or image level. Image-level fusion in this work refers to the fusion of information available through all the frames extracted from the videos. This can be considered as an example of combining information through multiple samples using a single sensor. This level of fusion occurs before matching is performed.

To evaluate the performance of image level fusion, the SR frames generated by all three techniques (SR3, SR5, AFS) are matched with gallery frames. The resulting scores are compared with the scores obtained by matching LR frames with the gallery frames. As template generation is not feasible using the LR frame set, the reference set is used to provide a benchmark.

5.5.2 Score-Level Fusion

In score-level fusion, the match score outputs of multiple biometric matchers can be fused to generate a new match score. Such fused scores can be used to make an identity decision in a biometric module. Score-level fusion is easy to perform when compared to image-level fusion. In this work, score-level fusion can be considered as the fusion of

match scores corresponding to the reference frames which are used for obtaining an SR frame.

The *sum rule* for score-level fusion is used in this work. If frames $\{f_1, f_2, ..., f_n\}$ are used to obtain an SR frame, the scores $\{S_1, S_2, ..., S_n\}$ obtained by matching the n frames with gallery frames are fused together by the sum rule, which is given by equation (5.4).

$$S_{sum} = \sum_{i=1}^{n} S_i. \tag{5.4}$$

5.6 Results for IIT-NRC Database

From the extracted LR frames, SR frames are generated using the three different techniques studied in this thesis: SR3, SR5, and AFS. The number of frames generated by SR3 and SR5 techniques are 6902 and 6858, respectively. The number of frames obtained by the AFS process is 1427. As the AFS technique only considers those frames which have a value of the motion parameter below a fixed threshold, the number of frames generated by this technique is fewer than the other two methods.

5.6.1 Image-Level Fusion

The ROC curves obtained by matching the SR frames generated by each process with the gallery frames are shown in Figure 5.3. The ROC curves and the Equal Error Rates (EER) reveal that the matching performance is high for the AFS technique when compared to SR3 and SR5 techniques. This supports the initial hypothesis that the elimination of frames degraded due to motion would lead to frames which contain maximum information. The presence of artifacts in SR3 and SR5 lowers the performance of the two techniques. As the number of artifacts occurring in SR3 is lower, its performance is higher than SR5.

It is noticed that apart from the performance gain, the computational cost is greatly reduced by employing AFS technique for super-resolving frames. AFS technique outputs a better performance, by fusing comparatively fewer frames and negates the necessity of redundant fusion of successive frames.

Identification rates are calculated using all the frames, generated by the three techniques, for a given video. Figure 5.4 shows the identification rates for all three techniques, with a comparison provided by the same analysis performed on the reference set (bicubic

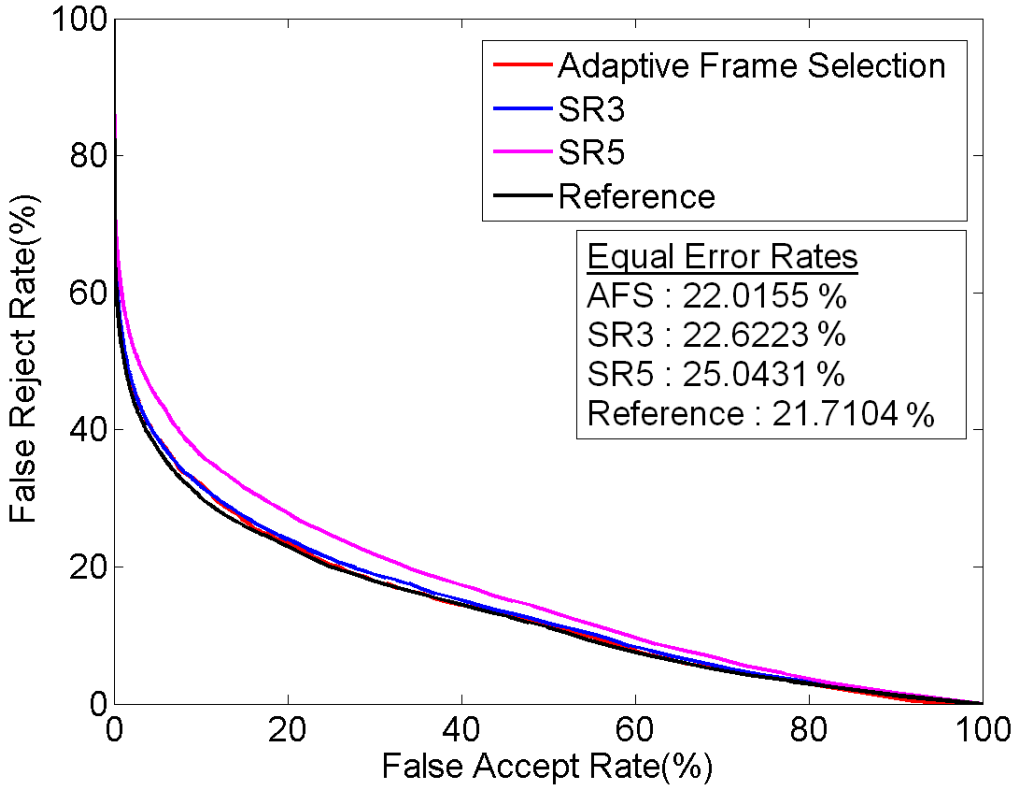

Figure 5.3: ROC curves of various processes for matching experiments.

interpolation).

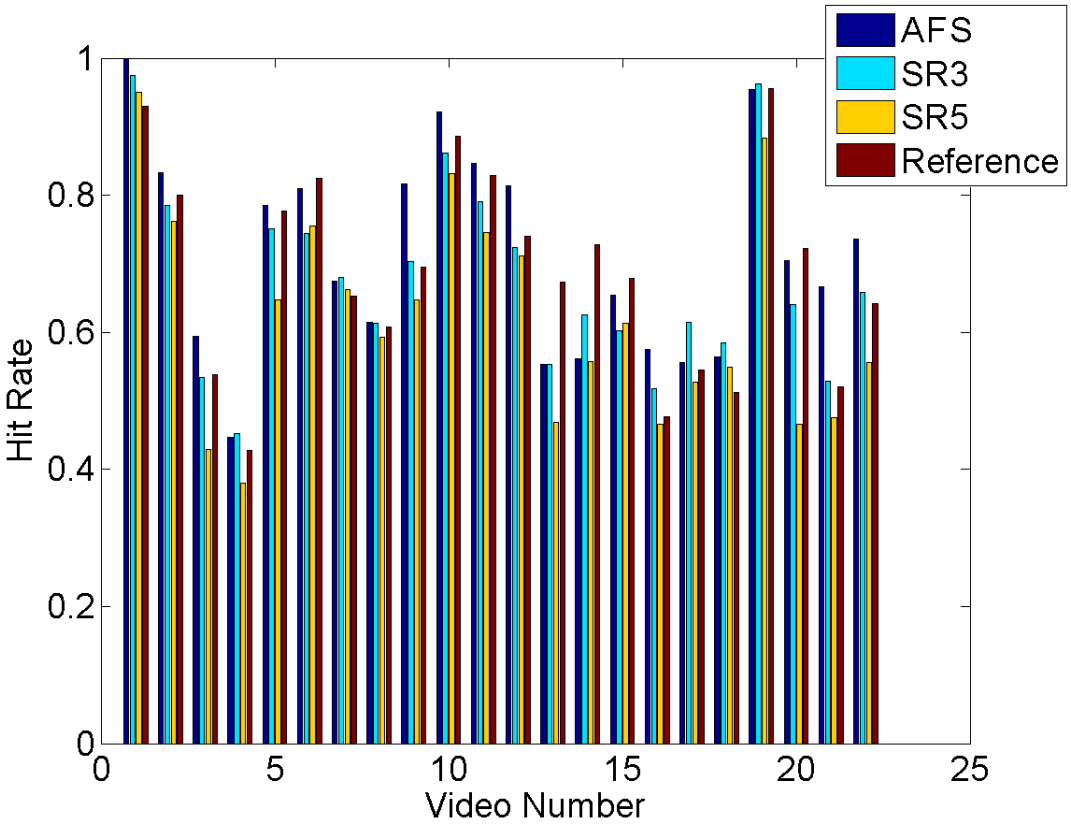

Figure 5.4: Identification rates for all the videos in the database.

The results show that AFS and bicubic interpolation techniques have better performances when compared to SR3 and SR5. For most videos, the performance of AFS is highest, with a comparable performance by the bicubic interpolation technique. As discussed earlier, the performances of SR3 and SR5 are affected by the presence of artifacts. Some of the frames reconstructed by these techniques produce faces in which the facial features are heavily degraded, reducing the identification performance.

To understand the variations in the identification rates across videos, it was necessary to observe an intrinsic property of a video which varied according to the ambient conditions present in the video. For this purpose, the mean value of β for a given video was observed. It is noticed that the identification rate for a given video for the AFS set is inversely related to the mean value of β for the frames for that video, as shown in Figure 5.5. This proves the effectiveness of β, supporting the hypothesis that if motion degradation occurs in a video, the identification task becomes challenging. The hit rates for AFS are higher than

the other techniques for videos whose β values are high. A large mean β value for a video indicates that the constituent frames contain large displacements or motion. Thus, from this experiment, it is inferred that when a video is heavily degraded by motion, the identification performance for such a video is improved by employing the AFS technique compared to the other techniques.

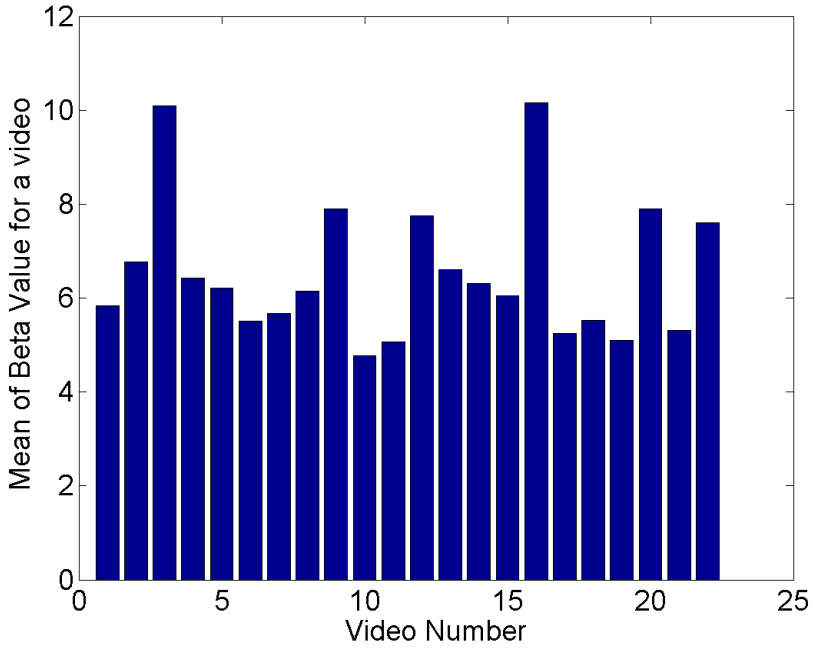

Figure 5.5: Mean values of β for a given video.

5.6.2 Score-Level Fusion

Score-level fusion is performed by fusing the match scores of all the reference frames corresponding to the LR frames used for image-level fusion. This procedure is repeated for all three techniques (SR3, SR5, AFS). Figure 5.6 shows the results obtained. The ROC curve of the bicubic interpolation is shown in the figure to provide a baseline for comparing the techniques.

From the results, it is observed that score-level fusion improves the performance of all three techniques compared to the corresponding image-level fusion schemes. Another crucial observation in the score-level fusion experiment is that the order of performances of the three techniques is reversed when compared to image-level fusion. The aforementioned

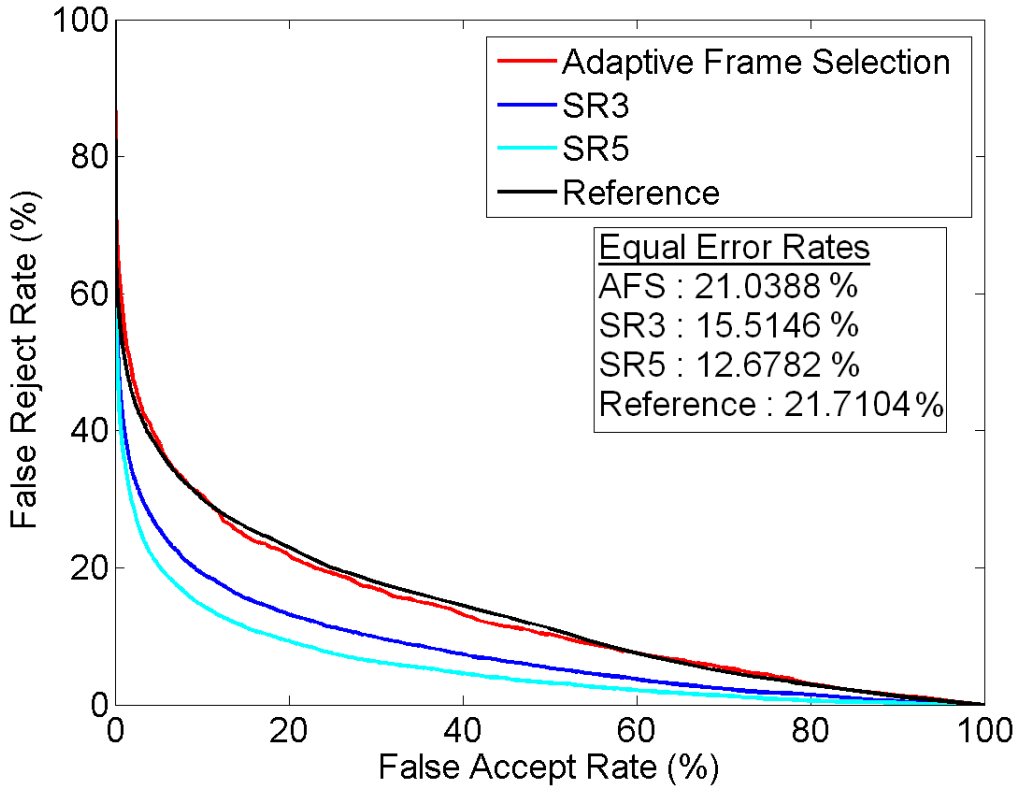

Figure 5.6: ROC curves using the score-level fusion for the three techniques. The ROC curve of the bicubic interpolation technique, is used for reference purposes.

observations could be related to the relative magnitude of match scores corresponding to the reference frames used for fusion. Even if only a subset of frames used for fusion have high match score values, this would heavily impact the final fused score. Such an impact could be large enough to negate the performance degradation due to super-resolution artifacts.

In the case of AFS, the component match scores corresponding to individual frames will be comparable, and thus do not yield higher levels of performance compared to that of SR3 and SR5.

5.7 Need for a different database

To compare the results of super-resolution, it is necessary to have a baseline. Most of the work on super-resolution in the literature report the performance by using a high resolution database. The high-resolution frames are first downsampled and then the downsampled frames are super-resolved. This gives a reference set of frames (the high-resolution images) which can be used to evaluate the performance of the super-resolved frames. In the case of the IIT-NRC database, it was not possible to perform such down-sampling due to the poor resolution of the frames to begin with.

For this purpose, a set of videos were collected at West Virginia University (WVU), under a controlled environment. A set of seven videos, one recording for each of seven different individuals, were obtained using a Logitech webcam, captured at a rate of 15 frames per second and a spatial resolution of 320×240 pixels. To maintain uniformity in movement across all the individuals, a protocol was designed which required the individual to narrate their name, and then move swiftly toward the left and right directions within a short period of time. This helps in capturing small and large motion displacements in a single video.

Since the WVU database allows the calculation of bivariate measures, quality assessment results are also reported on this database.

5.8 Results on the WVU Database

More than 850 frames were extracted from the videos, each having a resolution of 320×240, to form the reference set. An averaging process was used to convert the frames to a 160×120 pixel resolution. Super-resolution processes were applied on this low-resolution set to obtain the SR3, SR5, and AFS sets.

The number of LR frames was 859, with the number of frames generated by SR3 and SR5 techniques being 845 and 831, respectively. The number of frames generated by the AFS technique was 139. As explained earlier, since the AFS technique selects only those frames having small β values, the number of frames generated by this technique is low.

5.8.1 Image Level Fusion

Figure 5.7 shows the ROC curves for all three techniques with a baseline provided by the reference frame set. Results show that the order of performances of the three techniques are the same as that of the IIT-NRC database. Two significant observations are made regarding the performance of the AFS technique in this particular evaluation, which are listed below.

The performance of the AFS technique is very high when compared to the SR3 and SR5 techniques. In the IIT-NRC database, the performance of AFS was comparable to SR3 and SR5, but the ROC curves were not largely separated from each other. The reason for this improved performance in the WVU database is related with the inter-pupillary distances of the individuals. The inter-pupillary distance remains constant throughout the video in the WVU database whereas it is not constant in the IIT-NRC database. This reduces the performance of the individual techniques in the IIT-NRC database. Also, the videos in the IIT-NRC database had other conditions affecting the performance like multiple poses, varying expressions, etc. Figure 5.8 shows the variation of the inter-pupillary distance in the IIT-NRC database and Figure 5.9 shows a sample set of frames which depict the constancy of the inter-pupillary distance in the WVU database.

It is noticed that the Equal Error Rate for the AFS technique is zero. Figures 5.10, 5.11, and 5.12 show the distributions of genuine and impostor scores for various techniques. For the AFS technique, it is seen that the distributions are clearly separated. In the cases of SR3 and SR5, an overlap between the scores is noticed. The high perfor-

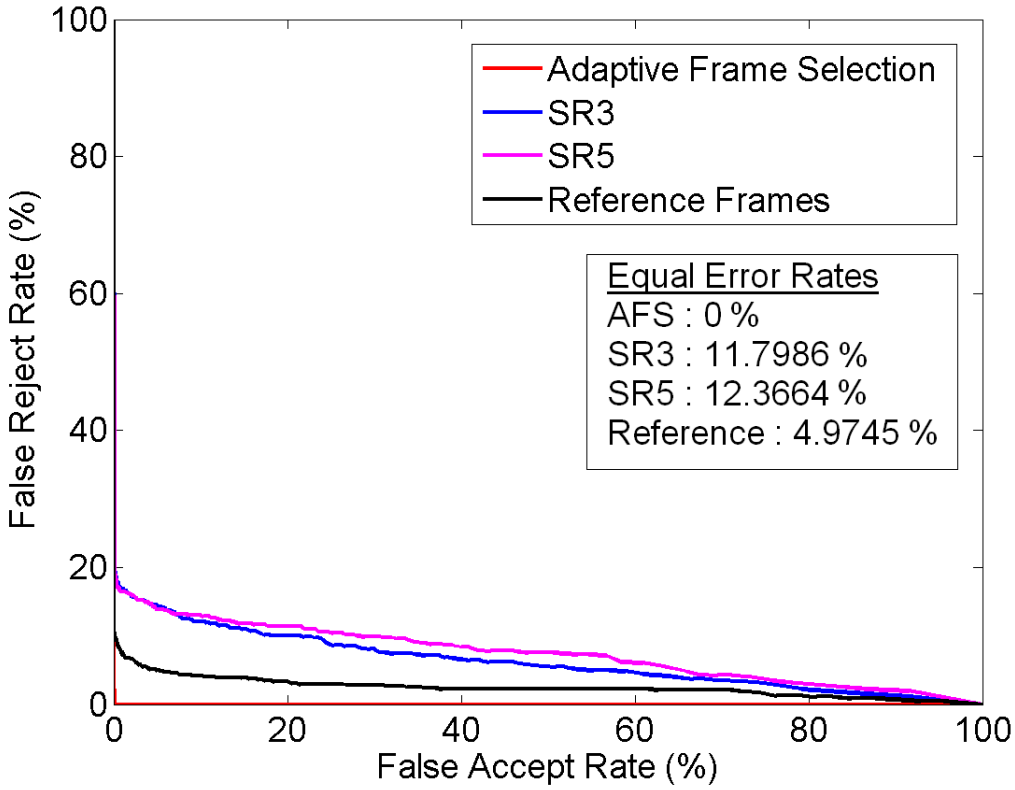

Figure 5.7: ROC curves for various techniques using the WVU database.

Figure 5.8: Variation in inter-pupillary distance in the IIT-NRC database.

Figure 5.9: Constancy in inter-pupillary distance in the WVU database.

mance of the AFS technique in this database can be attributed to the small size of the database.

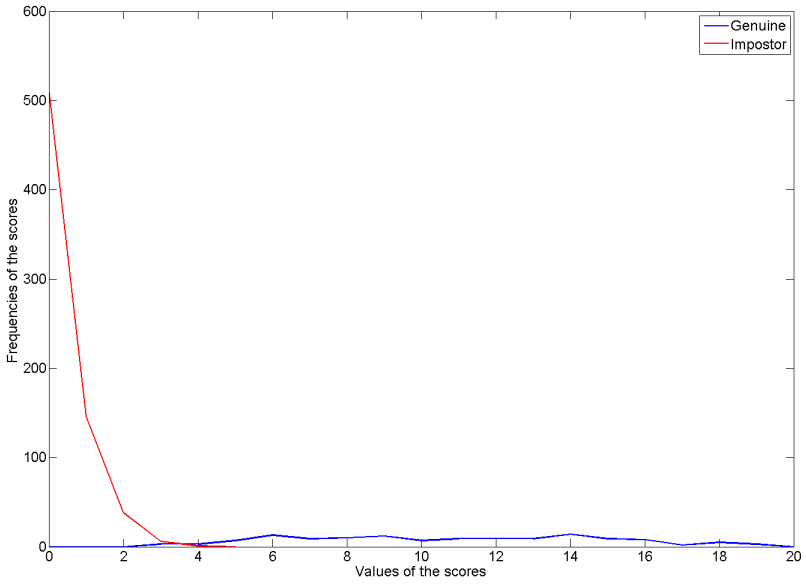

Figure 5.10: Distribution of genuine and impostor scores generated by using the AFS technique.

To justify the high performance of the AFS technique, ROC curves were plotted for two cases based on the frames used for fusion:

(a) frame set selected by the AFS technique, and

(b) the set of reference frames manually selected by a human corresponding to less motion displacements.

Figure 5.13 shows the ROC plots obtained for both the sets. The results show that the performances of both cases are comparable and thus proves the efficiency of the AFS technique in selecting frames with less motion.

Figure 5.14 shows the hit-rates plotted for every video in the WVU database and Figure 5.15 shows the mean values of β for the corresponding video. Results indicate that the hit-rates are inversely related to the mean values of β. This proves the earlier stated inference that whenever the mean value of β is high, the identification rate of the AFS technique is higher than the other two techniques. This also supports the earlier

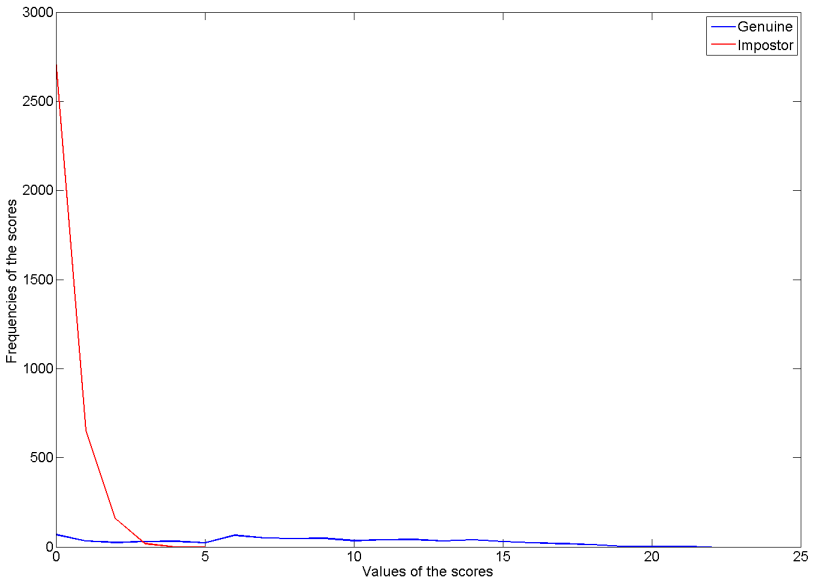

Figure 5.11: Distribution of genuine and impostor scores generated by using the SR3 technique.

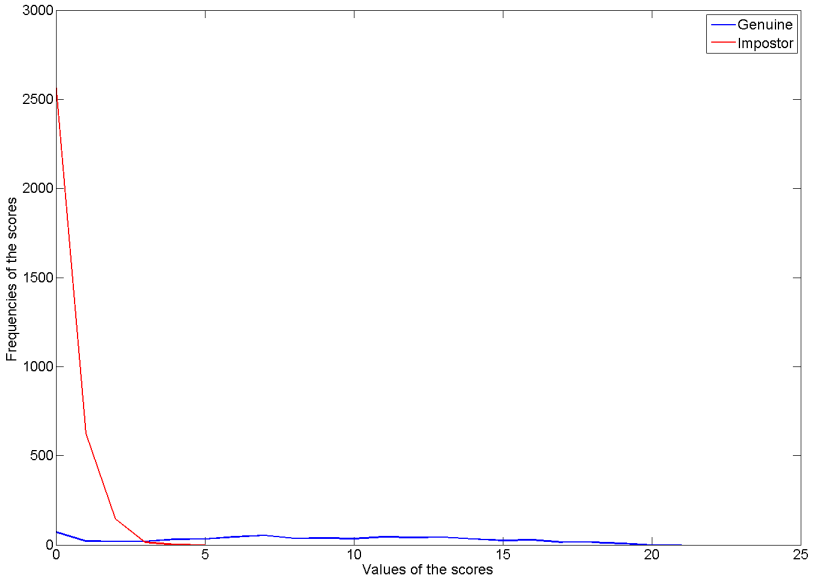

Figure 5.12: Distribution of genuine and impostor scores generated by using the SR5 technique.

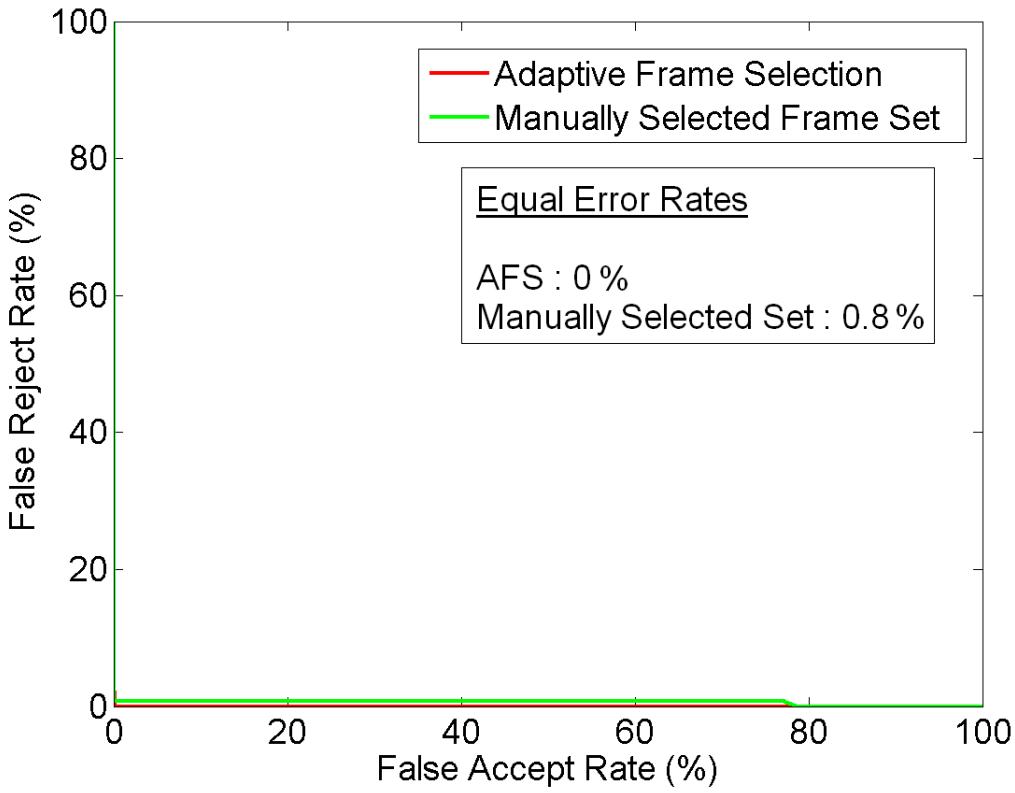

Figure 5.13: ROC curves indicating the performance of image level fusion when (a) AFS is used to select the frames, and (b) the frames manually selected.

stated observation that the using the AFS technique yields the best performance when the motion component is high in a video. Thus, it can be stated that the AFS technique is very suitable for super-resolution applications in surveillance video.

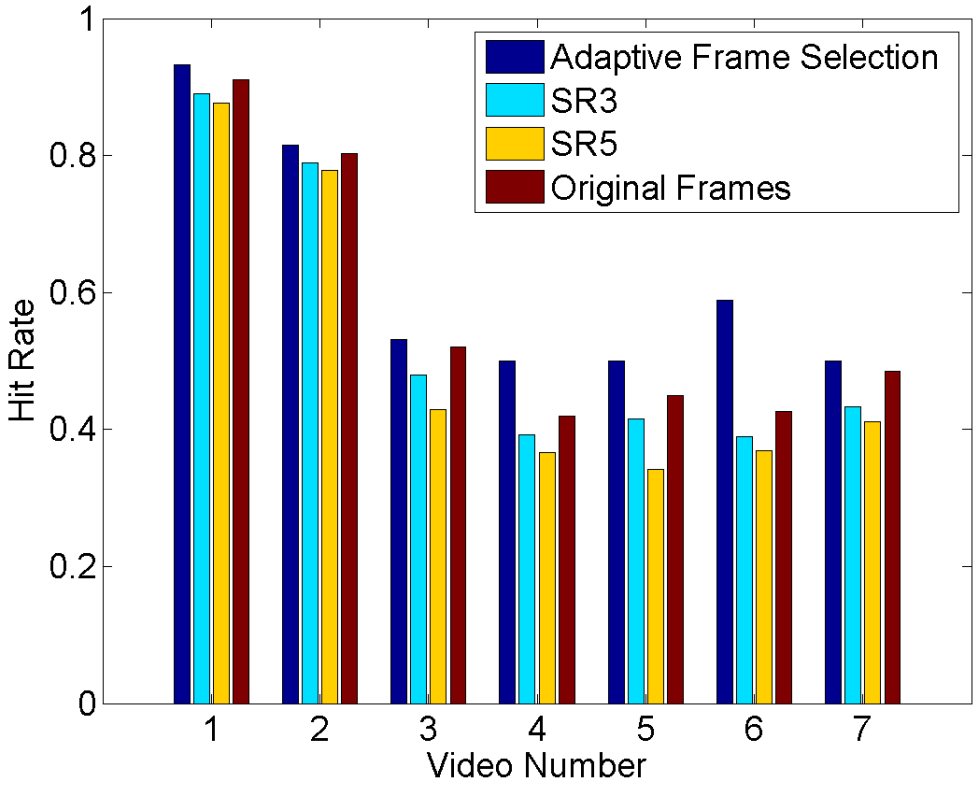

Figure 5.14: Identification rates for each video in the WVU database.

For evaluating the three techniques using quality metrics, the MSE values are used. A given super-resolved frame is compared with its reference frame to generate the MSE value. Only those super-resolved frames which correspond to the frame set generated by AFS technique are considered from the SR3 and SR5 sets.

After obtaining the MSE values, the difference between the MSE values of AFS and SR3, and AFS and SR5 are plotted. If a point in the difference plots lies below the horizontal axis, it indicates that the AFS technique results in a frame of better quality. Figures 5.16 and 5.17 show that the quality of images reconstructed using the AFS technique is generally higher compared to the SR3 and SR5 techniques. It is to be noted that, although the process used to reconstruct the frames are the same in all three techniques,

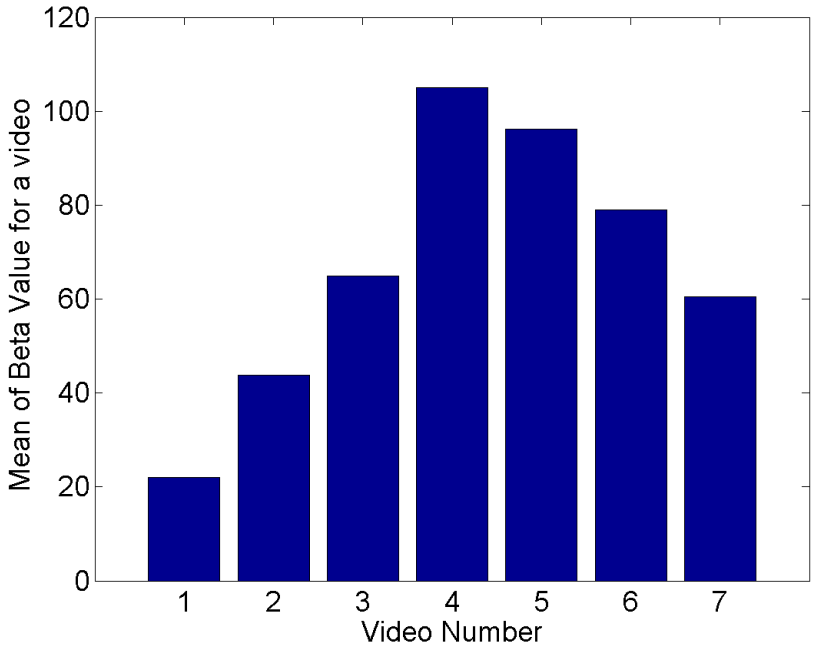

Figure 5.15: Mean values of β for individual videos in the WVU database.

the improvement in quality using AFS is due to the rejection of frames and the usage of appropriate number of frames resulting in minimum artifacts.

5.8.2 Score-Level Fusion

The ROC curves for score-level fusion in the WVU database are shown in Figure 5.18. It is again noted that the performance of the three techniques are improved compared to that of image-level fusion. This suggests the possibility of using score-level fusion to improve matching performance while reducing computation time compared to image level fusion.

Another observation made in this case is that the order of performances of the three techniques is the same as the image-level fusion. However, this was not the case for the IIT-NRC database. This is caused by two factors: first, the smaller number of identities in the WVU database, and second, the relatively less variation of the inter-pupillary distances in the WVU database.

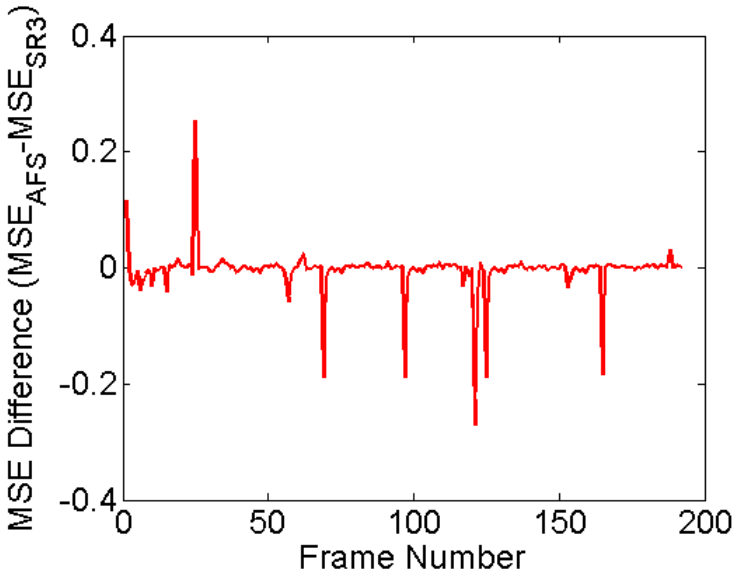

Figure 5.16: Difference in MSE values between AFS and SR3 techniques.

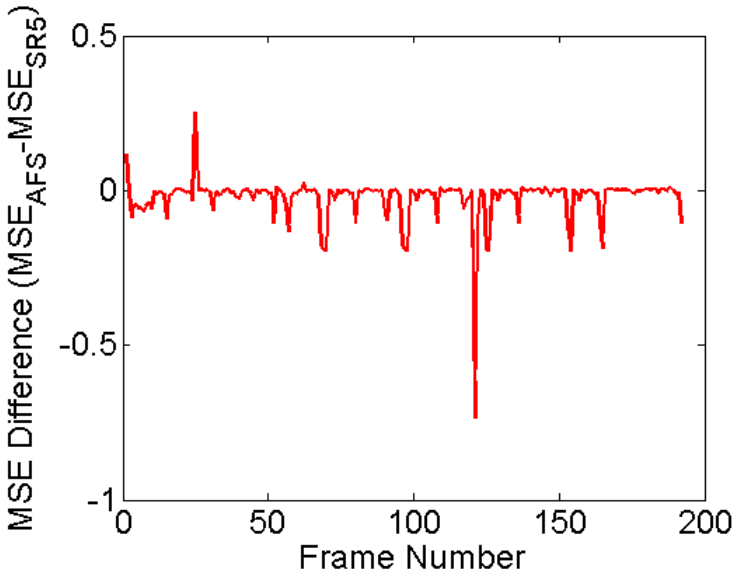

Figure 5.17: Difference in MSE values between AFS and SR5 techniques.

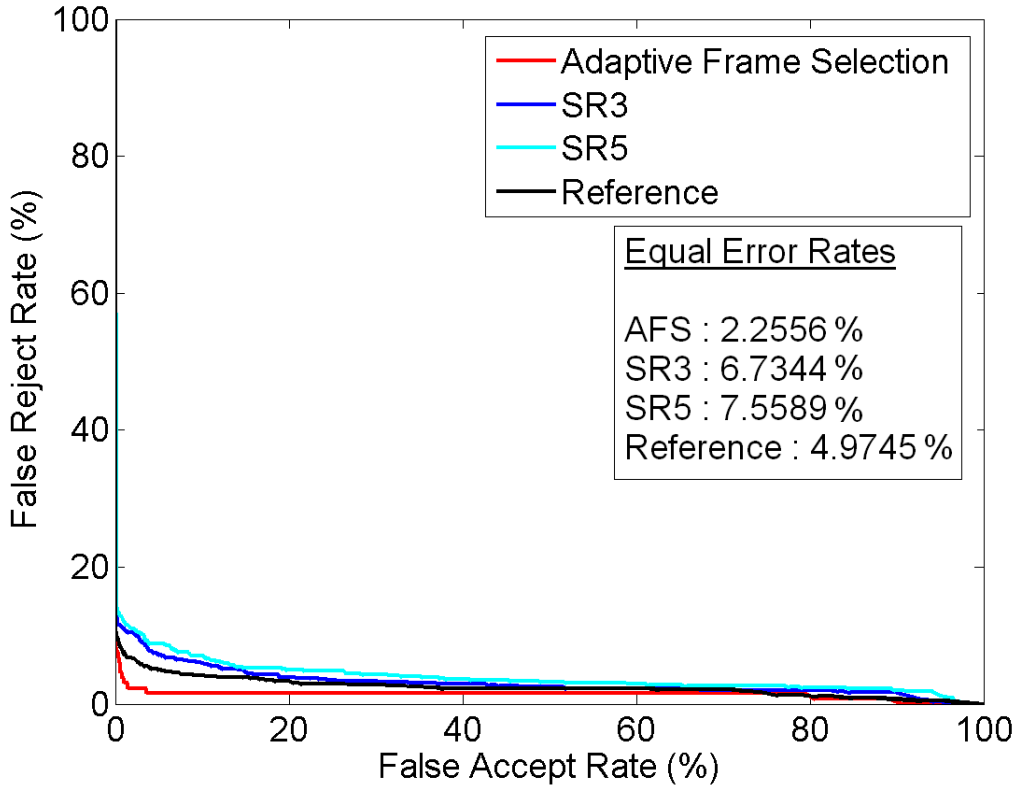

Figure 5.18: ROC curves for the score-level fusion scheme in the WVU database.

Chapter 6

Thesis Contributions and Future Work

6.1 Thesis Contributions

Super-resolution is an important image processing technique to extract maximum information from a low-resolution image. This thesis explored the applicability of super-resolution techniques exclusively for manipulating facial information from low-resolution surveillance videos.

The super-resolution optical flow technique, which can be effectively used for human faces in low resolution videos, was found to be a good candidate for super-resolving images in the above defined problem domain. However, a practical implementation of the technique on a low-resolution video database suggested that degradation due to large motion was a significant drawback.

It was desired to develop a technique that could eliminate the motion degradation effect, which is of crucial importance in surveillance videos. An adaptive frame selection technique was proposed in this thesis, which proves that the artifacts occurring due to motion degradation can be successfully eliminated by assessing the motion contained in successive frames and eliminating certain frames. It was also observed that due to the effective elimination of frames which would cause a decrease in performance, the quality of the reconstructed frame is improved.

The experiments in the thesis also support the applicability of adaptive frame selection

for identification purposes, especially in cases where motion plays a significant role. Since the adaptive frame selection technique results in better identification rates, the usage of this technique for low-resolution surveillance videos is recommended.

Another benefit of adopting this work lies in the fact that the frames which are eliminated typically correspond to large changes in scene or pose. This could help in automatically selecting templates with high intra-class variations. Other important advantages of the adaptive frame selection technique are that it is computationally inexpensive and less time consuming compared to other techniques.

It was noticed that the quality of the output image generated by adaptive fusion technique is better than that of the other two considered techniques. This thesis presents a prefatory effort towards mathematically formulating motion between two successive frames using optical flow matrices. Effective calculations of motion is of major significance in various image processing tasks. Apart from the adaptive technique to fuse frames, this thesis also compares the performance of image-level and score-level fusion schemes in the realm of low resolution facial images. Results show that score-level fusion performs better than image-level fusion.

6.2 Future Work

The effect of threshold selection for selecting frames has to be studied in detail. Also, the effectiveness of the value of β needs to be computed by comparing the quality of frames reconstructed by using a particular value of β with that of the visually assessed quality of the frames.

Various mathematical techniques could be explored for calculation of β after the optical flow between images is computed. Also, results can be reported using various other databases and then performances could be compared to support the current inferences. Work could be carried out to estimate 3D motion such as yaw, pitch or roll occurring between successive frames, based on optical flow matrices.

Also, apart from the sum rule for score-level fusion the effect of using other fusion rules like min rule, max rule etc. can be studied.

References

[1] Multiple Biometric Grand Challenge, "http://face.nist.gov/mbgc/," .

[2] A.M. Martinez and R. Benavente, "The AR face database," CVC Technical Report 24, Purdue University, June 1998.

[3] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," *7th International Conference on Automatic Face and Gesture Recognition*, pp. 341–345, April 2006.

[4] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167 – 1183, September 2002.

[5] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[6] A. Pentland and T. Choudhury, "Face recognition for smart environments," *Computer*, vol. 33, no. 2, pp. 50–55, February 2000.

[7] S. Pankanti, R. M. Bolle, and A. Jain, "Biometrics-the future of identification," *Computer*, vol. 33, no. 2, pp. 46–49, 2000.

[8] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, Jin Chang, K. Hoffman, J. Marques, M. Jaesik, and W. Worek, "Overview of the face recognition grand challenge," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR)*, vol. 1, pp. 947–954, June 2005.

[9] K. Mikolajczyk, R. Choudhury, and C. Schmid, "Face detection in a video sequence - a temporal approach," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. II–96–II–101 vol.2, 2001.

[10] G.L. Foresti, C. Micheloni, L. Snidaro, and C. Marchiol, "Face detection for visual surveillance," *Proceedings of the 12th International Conference on Image Analysis and Processing*, pp. 115–120, September 2003.

[11] A. Destrero, F. Odone, and A. Verri, "A trainable system for face detection in unconstrained environments," in *Proceedings of the 14th International Conference on Image Analysis and Processing (ICIAP)*, Washington, DC, USA, 2007, pp. 407–412, IEEE Computer Society.

[12] M. Castrillón, O. Déniz, C. Guerra, and M. Hernández, "Encara2: Real-time detection of multiple faces at different resolutions in video streams," *Journal of Visual Communication and Image Representation*, vol. 18, no. 2, pp. 130–140, 2007.

[13] G.L. Foresti, P. Mehanen, and C.S. Regazzoni, *Multimedia Video-Based Surveillance Systems*, Kluwer, September 2000.

[14] S. Chaudhri, *Super-Resolution Imaging*, Kluwer Academic, Boston, Mass, USA, 2001.

[15] J. D. van Ouwerkerk, "Image super-resolution survey," *Image and Vision Computing*, vol. 24, no. 10, pp. 1039–1052, 2006.

[16] F. Lin, C. Fookes, V. Chandran, and S. Sridharan, "Super-resolved faces for improved face recognition from surveillance video," in *ICB*, S. W. Lee and S. Z. Li, Eds. 2007, vol. 4642 of *Lecture Notes in Computer Science*, pp. 1–10, Springer.

[17] R. Y. Tsai and T. S. Huang, "Multiframe image restoration and registration," *Advances in Computer Vision and Image Processing*, vol. 1, pp. 317–339, 1984.

[18] S. Borman and R.L. Stevenson, "Super-resolution from image sequences-a review," *Proceedings. of the Midwest Symposium on Circuits and Systems*, pp. 374–378, August 1998.

[19] G. Wolberg, *Digital Image Warping*, IEEE Computer Society Press, Los Alamitos, CA, USA, 1994.

[20] W. K. Pratt, *Digital Image Processing (2nd ed.)*, John Wiley & Sons, Inc., New York, NY, USA, 1991.

[21] K. Xue, A. Winans, and E. Walowit, "An edge-restricted spatial interpolation algorithm," *Journal of Electronic Imaging*, vol. 1, no. 2, pp. 152–161, 1992.

[22] N. B. Karayiannis and A. N. Venetsanopoulos, "Image interpolation based on variational principles," *Signal Processessing*, vol. 25, no. 3, pp. 259–288, 1991.

[23] R. R. Schultz and R. L. Stevenson, "A bayesian approach to image expansion for improved definition," *IEEE Transactions on Image Processing*, vol. 3, no. 3, pp. 233–242, May 1994.

[24] R.R. Schultz and R.L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Transactions on Image Processing*, vol. 5, no. 6, pp. 996–1011, June 1996.

[25] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *International Journal of Computer Vision*, vol. 40, pp. 2000, 2000.

[26] S. Baker and T. Kanade, "Hallucinating faces," Tech. Rep. CMU-RI-TR-99-32, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, September 1999.

[27] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[28] M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1646–1658, December 1997.

[29] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *In ECCV*. 1992, pp. 237–252, Springer-Verlag.

[30] M. Elad and A. Feuer, "Super-resolution restoration of an image sequence: adaptive filtering approach," *IEEE Transactions on Image Processing*, vol. 8, pp. 387–395, 1999.

[31] S. P. Kim, N. K. Bose, and H. M. Valenzuela, "Recursive reconstruction of high resolution image from noisy undersampled multiframes," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 6, pp. 1013–1027, June 1990.

[32] C. Srinivas and M. D. Srinath, "A stochastic model based approach for simultaneous restoration of multiple misregistered images," *Proceedings of SPIE*, vol. 1360, pp. 1416–1427, 1990.

[33] A.M. Tekalp, M.K. Ozkan, and M.I. Sezan, "High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 169–172, 1992.

[34] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *Journal of the Optical Society of America A*, vol. 6, no. 11, pp. 1715–1726, 1989.

[35] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical Model and Image Processing*, vol. 53, no. 3, pp. 231–239, 1991.

[36] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint map registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1621–1633, 1997.

[37] S. Baker and T. Kanade, "Super resolution optical flow," Tech. Rep. CMU-RI-TR-99-36, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, October 1999.

[38] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, 1994, pp. 593 – 600.

[39] J. Y. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker," Intel Corporation, Microprocessor Research Labs, http://www.intel.com/research/mrl/research/opencv/, 2000.

[40] F. L. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[41] L. G. Brown, "A survey of image registration techniques," *ACM Computing Surveys*, vol. 24, pp. 325–376, 1992.

[42] D. I. Barnea and H. F. Silverman, "A class of algorithms for fast digital imiage registration," *IEEE Transactions on Computers*, vol. C-21, no. 2, pp. 179–186, 1972.

[43] T. Acharya and P. S. Tsai, "Computational foundations of image interpolation algorithms," *Ubiquity*, vol. 8, no. 42, pp. 1–17, 2007.

[44] S. Thurnhofer and S. K. Mitra, "Edge-enhanced image zooming," *Optical Engineering*, vol. 35, no. 7, pp. 1862–1870, 1996.

[45] A. J. Parker, R. V. Kenyon, and D. E. Troxel, "Comparison of interpolating methods for image resampling," *IEEE Transactions on Medical Imaging*, vol. 2, no. 1, pp. 31–39, March 1983.

[46] T. M. Lehmann, C. Gonner, and K. Spitzer, "Survey: Interpolation methods in medical image processing," *IEEE Transactions on Medical Imaging*, vol. 18, no. 11, pp. 1049–1075, 1999.

[47] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001.

[48] A. K. Katsaggelos, *Digital Image Restoration*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1991.

[49] A. K. Katsaggelos, "Iterative image restoration algorithms," *Optical Engineering*, vol. 28, no. 7, pp. 735–748, July 1989.

[50] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Computing Surveys*, vol. 27, no. 3, pp. 433–467, 1995.

[51] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[52] F. Bartolini, V. Cappellini, Colombo C., and A. Mecocci, "Enhancement of local optic flow techniques," *Proceedings 4th International Workshop on Time Varying Image Processing and Moving Object Recognition, Florence, Italy*, pp. 359–366, June 10-11 1993.

[53] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, April 1981, pp. 674–679.

[54] M. C. Chiang and T. E. Boult, "Efficient image warping and super-resolution.," in *In IEEE Workshop on Applications of Computer Vision (WACV)*. 1996, pp. 56–61, IEEE Computer Society.

[55] F. Lin, J. Cook, V. Chandran, and S. Sridharan, "Face recognition from super-resolved images," *Proceedings of the Eighth International Symposium on Signal Processing and Its Applications*, vol. 2, pp. 667–670, 28-31, 2005.

[56] F. Lin, J. Cook, V. Chandran, and S. Sridharan, "Investigation into optical flow super resolution for surveillance applications," *Proceedings of APRS Workshop on Digital Image Computing*, pp. 73–78, February 2005.

[57] R. de Freitas Zampolo and R. Seara, "A measure for perceptual image quality assessment," *Proceedings of the International Conference on Image Processing (ICIP)*, vol. 1, pp. I–433–6 vol.1, September 2003.

[58] A. M. Eskicioglu and P.S. Fisher, "A survey of quality measures for gray scale image compression," *Space and Earth Science Data Compression Workshop*, pp. 49–61, April 1993.

[59] K. Hosaka, "A new picture quality evaluation method," *Proceedings of International Picture Coding Symposium, Tokyo, Japan*, pp. 17–18, April 1986.

[60] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on Communications*, vol. 43, no. 12, December 1995.

[61] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing using MATLAB*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003.

[62] I. Avcibas, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," *Journal of Electronic Imaging*, vol. 11, pp. 206–223, 2002.

[63] Z. Wang and A.C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, Mar 2002.

[64] X. Li, "Blind image quality assessment," *Proceedings of International Conference on Image Processing.*, vol. 1, pp. I–449–I–452 vol.1, 2002.

[65] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *Proceedings of IEEE International Conferencing on Image Processing*, 2002, pp. 22–25.

[66] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, 2004.

[67] R. L. V. Hsu, J. Shah, and B. Martin, "Quality assessment of facial images," *Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*, pp. 1–6, August 2006.

[68] D. O. Gorodnichy, "Video-based framework for face recognition in video," in *CRV '05: Proceedings of the 2nd Canadian conference on Computer and Robot Vision*, Washington, DC, USA, 2005, pp. 330–338, IEEE Computer Society.

[69] F. Oberti, A. Teschioni, and C.S. Regazzoni, "ROC curves for performance evaluation of video sequences processing systems for surveillance applications," *Proceedings of the International Conference on Image Processing, (ICIP)*, vol. 2, pp. 949–953 vol.2, 1999.

[70] FaceIt, "SDK developer's guide," Software Version 6.1, 2005.

[71] A. Ross, "An introduction to multibiometrics," *Proceedings of the 15th European Signal Processing Conference (EUSIPCO), Poznan, Poland*, September 2007.