5-2013

Motion Segmentation Aided Super Resolution Image Reconstruction

Muharrem Mercimek
mmercime@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Muharrem Mercimek entitled "Motion Segmentation Aided Super Resolution Image Reconstruction." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Electrical Engineering.

Mongi A. Abidi, Major Professor

We have read this dissertation and recommend its acceptance:

Andreas Koschan, Seddik M. Djouadi, Hamparsum Bozdogan

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Motion Segmentation Aided Super Resolution Image Reconstruction

A Dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Muharrem Mercimek
May 2013

This dissertation is dedicated to:

My family Hatice Mercimek, Nedim Mercimek,
My mother İclal Mercimek, My father Nedim Mercimek,
My sisters Meltem Bilge and Melike Mercimek
My supportive friends, and instructors,
for always believing in me, inspiring me, and encouraging me
to reach higher in order to achieve my goals.

Acknowledgements

I am deeply thankful to my parents, Nedim Mercimek and İclal Mercimek who have always supported me, to my dear wife Hatice Mercimek for her understanding, patience and encouragement during my life and education, and to my son Nedim Mercimek, who is the sunshine of my life.

I would like to thank my advisor Dr. Mongi A. Abidi who gave me an opportunity to work under his guidance. Without his support and encouragement, this work would not have been possible. I sincerely thank Dr. Andreas Koschan whose technical comments provided me valuable guidance throughout my research. Special thanks goes to my committee members: Dr. Seddik M. Djuadi, Dr. Hamparsum Bozdogan for their valuable advices and encouragement. I greatly appreciate their time and input to this dissertation.

I would also like to thank to; all IRIS members for their assistance during my studies.

Abstract

This dissertation addresses Super Resolution (SR) Image Reconstruction focusing on motion segmentation. The main thrust is Information Complexity guided Gaussian Mixture Models (GMMs) for Statistical Background Modeling. In the process of developing our framework we also focus on two other topics; motion trajectories estimation toward global and local scene change detections and image reconstruction to have high resolution (HR) representations of the moving regions. Such a framework is used for dynamic scene understanding and recognition of individuals and threats with the help of the image sequences recorded with either stationary or non-stationary camera systems.

We introduce a new technique called Information Complexity guided Statistical Background Modeling. Thus, we successfully employ GMMs, which are optimal with respect to information complexity criteria. Moving objects are segmented out through background subtraction which utilizes the computed background model. This technique produces superior results to competing background modeling strategies.

The state-of-the-art SR Image Reconstruction studies combine the information from a set of unremarkably different low resolution (LR) images of static scene to construct an HR representation. The crucial challenge not handled in these studies is accumulating the corresponding information from highly displaced moving objects. In this aspect, a framework of SR Image Reconstruction of the moving objects with such high level of displacements is developed. Our assumption is that LR images are different from each other due to local motion of the objects and the global motion of the scene imposed by non-stationary imaging system. Contrary to traditional SR approaches, we employed several steps. These steps are; the suppression of the global motion, motion segmentation accompanied by background subtraction to extract moving objects, suppression of the local motion of the segmented out regions, and super-resolving accumulated information coming from moving objects rather than the whole scene. This results in a reliable offline SR Image Reconstruction tool which handles several types of dynamic scene changes, compensates the impacts of camera systems, and provides data redundancy through removing the background. The framework proved to be superior to the state-of-the-art algorithms which put no significant effort toward dynamic scene representation of non-stationary camera systems.

Table of Contents

List of Tables

List of Figures

1 Introduction

In many fields ranging from security to medicine, a need has been driven for better understanding of a scene, especially to extract regions of interest. An increase in the sampling rate could be achieved by obtaining more information about scene from a sequence of images. The idea behind Super Resolution (SR) Image Reconstruction is to combine the complementary information from a set of different low resolution (LR) images (Figure 1.1) of the underlying scene and use it to construct a high resolution (HR) still image or a video, which is a better representation of the scene with more resolving power [Vandawalle06].

Resolution is a widely used term when judging various image acquisition/processing systems' quality, and it is mostly related to the sensor characteristics; density and spatial response of the detector elements. Among several of digital image resolution definitions, spatial resolution is commonly meant, referring to the number of independent pixel values per unit length. The smallest discernible and measurable detail in a visual presentation is also used as the definition of resolution in Optics.

Monitoring and video based systems has gained a big acceleration with the advances in electronics, sensors, and optics since the 1970's. Charge-coupled device (CCD), charge-injection device (CID) and complementary metal–oxide–semiconductor (CMOS) image sensors have been commonly used to capture digital images [Park03]. The increase in the number of the imaging sensors' elements clearly enhances the resolving power of the acquired images. However, this is not always practicable due to the increasing associated cost. Moreover, the shot noise increases during acquisition as the pixel size becomes smaller. One other definition of resolution; temporal resolution is the frame rate or the number of frames captured per second. The temporal resolution should be set proportional to the amount of motion in the image sequences. The tradeoff between temporal resolution of a spatio-temporal data and its spatial resolution is the bottleneck, in many image acquisition systems. One can favor having as many frames as possible and risk having high spatial resolution, which is often the case driven by daily applications. For surveillance systems, to record the scene for a long time the spatial resolution can be kept low on purpose. Some of the products in the market have capabilities of video streaming, which enable display of the frames over the web from any location [Katsaggelos07]. Due to bandwidth limitations, the frames have to be captured at a lower rate. In most medical imaging applications (Computed Tomography (CT) and X-ray scans) to align the magnetization of specific atoms in the body the powerful magnetic field is operated at a very low permitted level in a short time, accordingly the output frames are spatially poor. LR images are used out of necessity considering the high cost and physical limitations. All these key factors have a direct influence on sensor manufacturing techniques to be replaced with signal processing techniques to obtain HR images.

Considering that the resolution of HR image is higher than that of the LR frames, and the Nyquist sampling criterion is satisfied, this HR image is a more sufficient and better representation of the continuous scene. Thus, having frames containing fine details is satisfied. For one single observation the problem is ill-posed (Figure 1.2), since multiple HR images can be reduced to the same LR image. In other words, we have multiple ways to go from a single LR to an HR image.

The common method for such a problem is to constrain the solution space according to a priori knowledge of the form of solution [Borman98]. Having a number of interdependent images available adds stability to

SR Image/Video Reconstruction process. Throughout the study, we can use two terms image/video reconstruction or image reconstruction which are referring to the same process. LR images are the different looks at the same scene at different times. If the LR images are exposed to different sub-pixel level shifts naturally during image acquisition, then each image can provide supplementary information to obtain HR image. A possible example to set abundant looks to the scene is that the user is holding a digital camera taking a series of images in a very short time. The small vibrations of the user's hand during image acquisition are sufficient to reconstruct an HR image by SR techniques. This requires knowledge of the exact image displacements, which may happen to be available for the images acquired with experimentally controlled sub-pixel camera displacement. A similar scenario where image acquisition system is mounted on a mobile platform such as an aircraft or a robot, and observed objects are in the far field [Hardie97]. When we are having LR images, which are the synthetically shifted versions of each other by integer units, we will observe exactly the same information which makes the system not suitable for SR Image/Video Reconstruction. The complementary information in general is obtained naturally during image acquisition and cannot be imposed on a single image. Solution of the system requires that each observation contributes differently. From now on we will use the term SR interchangeably with SR Image/Video Reconstruction.

That the loss of high frequency components such as edges and textures produces distortions on LR images is a strong motive to employ SR, yet we do not focus on this aspect of SR in this study. When representing a scene with high levels of details if we cannot have dense set of pixels, the resulting image will suffer from aliasing artifacts. SR is not only useful to enhance the resolving power of an image in the imaging process; it can also, to some extent, reduce the aliasing noticeably [Vandawalle06]. Intrinsically, each LR image is a subsampled (i.e., aliased) version of the scene. The aliasing could not be removed if we were to process one image only. SR also drives the extrapolation of frequency content beyond that which is present in the observed data.

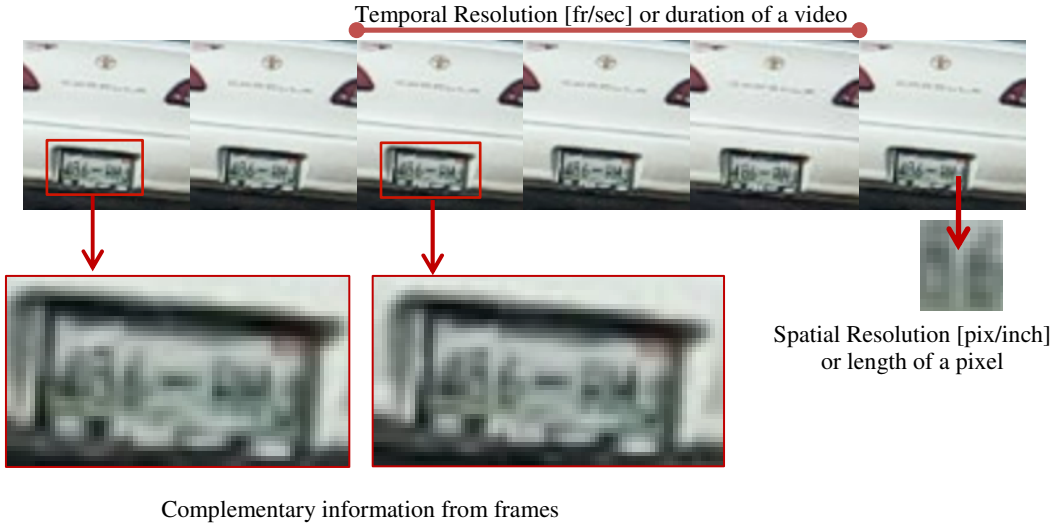


Figure 1.1: Multiple frames provide complementary information about the scene. There is a tradeoff between spatial resolution and temporal resolution of an image sequence.

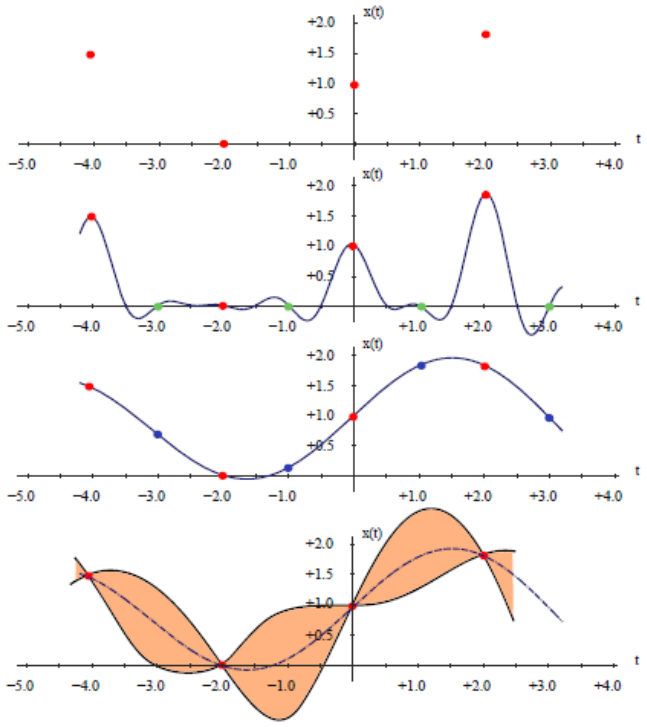


Figure 1.2: 1-D illustration of single frame interpolation ill-posedness; The LR data, two HR function candidates, set of possible HR function candidates obtained using given the boundaries of ambiguity.

In the literature SR is mostly utilized as a mean to ameliorate the undesirable reduction in resolution introduced by the imaging system [Katsaggelos07]. An illustration of SR Image Reconstruction in Practical use is given in Figure 1.3 also a few application areas are given in Figure 1.4. Accurate highlighting via zooming in is required in many image processing applications, such as in surveillance, forensic, satellite, and medical imaging. Reconstructing higher quality digital image from LR images obtained with an inexpensive camera/camcorder can be an application area. Recently, Close Circuit Television (CCTV) system has been replaced with Digital Video Recorder (DVR). Extraction of specific regions in the scene such as the face of a person or the license plate of a suspected vehicle for surveillance or forensic purposes may be needed. The studies on iris for personal identification and authentication have been conducted many times, and SR techniques can assist in having HR iris representation. In medical imaging such as CT and Magnetic Resonance Imaging (MRI) when searching abnormal activities, the resolution of multiple images is limited due to the radiation exposure concerns. SR techniques are resorted to overcome the resolution problems arising from the fact that the resolution must be kept at a certain level to provide a safe mean for the patients. Several images of the same area are usually provided in satellite imaging applications such as remote sensing. The SR techniques can improve the resolution of targets in that case. Indeed the first work on SR was aimed at improving the resolution of Landsat images by Tsai and Huang [Tsai84]. Another application is conversion from a National Television Signal Committee (NTSC) video signal to a High Definition Television (HDTV) signal. A capability to automatically identify and extract the contents of video would produce convenient indexed referencing which fits to a great number of video processing applications. Also, Optical Character Recognition (OCR) process will perform better when sufficiently super-resolved text images produced by SR algorithms are available.

Figure 1.3: SR Image Reconstruction in Practical Use [Sroubek07].

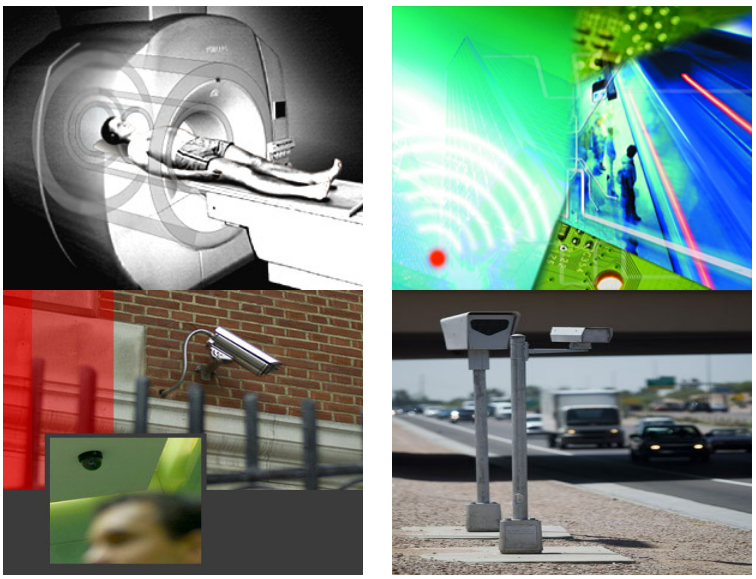


Figure 1.4: Some applications utilizing SR Image Reconstruction; Medical Imaging, Wireless Camera Networks, Building Surveillance, Roadwatch Traffic Surveillance.

The remainder of this chapter outlines the motivation for this research in section 1.1. Section 1.2 gives the pipeline of our system; Section 1.3 gives the contributions of our work. Section 1.4 concludes this chapter with the document organization.

1.1 Motivation

In SR literature, mostly the key idea is to get rid of sub-pixel misalignments imposed on a set of images by the use of different means. The images are aligned onto the same spatial coordinate system with sub-pixel accuracy in respect to rigid body transformations. This opens a pathway for image resolution enhancement. The assumption of having just slightly different LR images of the same scene to construct an HR image is mostly the case in these studies and some common origins such as noise, camera vibration, change of focus, or a combination of these impose variations on the scene.

The needs are dynamic scene understanding and recognition of individuals and threats with the help of image sequences. As we stated before LR images represent different looks at the same scene at different times. Our main assumption about the discrepancies observed in the scene representations is; for stationary camera system moving regions of the dynamic scene are related by local displacements, and for non-stationary camera system these regions are related by local displacements, additionally possible global displacements could be imposed on the whole scene.

In this aspect, LR image sequence containing moving objects with independent motion trajectories has to be dealt with. Global and local displacements should be estimated and recovered accurately to accumulate all the information related the corresponding regions from LR images. Main objective is to generate super-resolved representation of the moving objects rather than that of whole scene. Using the motion cues, motion areas has to be detected reliably. Due to areas of constant intensity values within the moving objects we do not receive dense motion vector fields. The identified regions of a moving object contain gaps and holes and moreover the aperture problem, which will be discussed in Chapter 3, cause parts of the objects to be left out. Thus, through utilizing an efficient background modeling the regions in motion from background which has no significant importance has to be distinguished more efficiently.

1.2 Block Diagram of Our Framework

Main efforts in this research are directed towards Information Complexity guided Gaussian Mixture Models (GMMs) for Statistical Background Modeling and the ultimate goal is image reconstruction to have HR representations of the extracted regions of interest. In the process of developing our framework we can put our efforts in two sets of processing blocks as shown in Figure 1.5; background subtraction or motion segmentation and image/video reconstruction.

For the background subtraction or motion segmentation block, basically there are five steps we are employing;

a. Pre-processing the data to stabilize the camera effects through estimation of the global motion trajectories imposed on the frames, due to effects such as wind load, and vibration, etc.
b. Initial background modeling and the maintenance using background models. The model should handle situations where the background of the scene is cluttered and also containing different types of motions. We discuss modeling each pixel using GMM and using an information complexity guided optimal GMM selection scheme, which is a new technique in background modeling field. This results in a stable moving target segmentation which reliably overcomes the demanding challenges of lighting changes, repetitive motions from clutter, and long-term scene changes.

c. Foreground detection also known as background subtraction, simply thresholding the difference between estimated model of the background, which is anticipated to contain no moving objects, and the current image.
d. Post-processing to obtain the final silhouette of the moving objects using multiple morphological operations and thresholding to suppress false detections that are due to small motions not captured by the model.
e. Rough local motion estimation of the segmented out moving object where the information related to the moving regions are localized.

In standard SR Image Reconstruction algorithms, the LR scene representations are introduced directly to SR Image Reconstruction block without considering moving regions in the scene. The answer to the question of why we need motion segmentation is illustrated in Figure 1.6. Having a set of different representations of the moving objects extracted from the background we can involve with the next step localization of all the information coming from moving objects. Apparently, rigid body transformation model is fine enough for our purposes. We will carry out our discussion on motion in Chapter 3. Overall main objective of motion estimation is that we want to force the corresponding moving regions of multiple images to be tightly close to each other in order to use abundant information efficiently in the SR Image Reconstruction block.

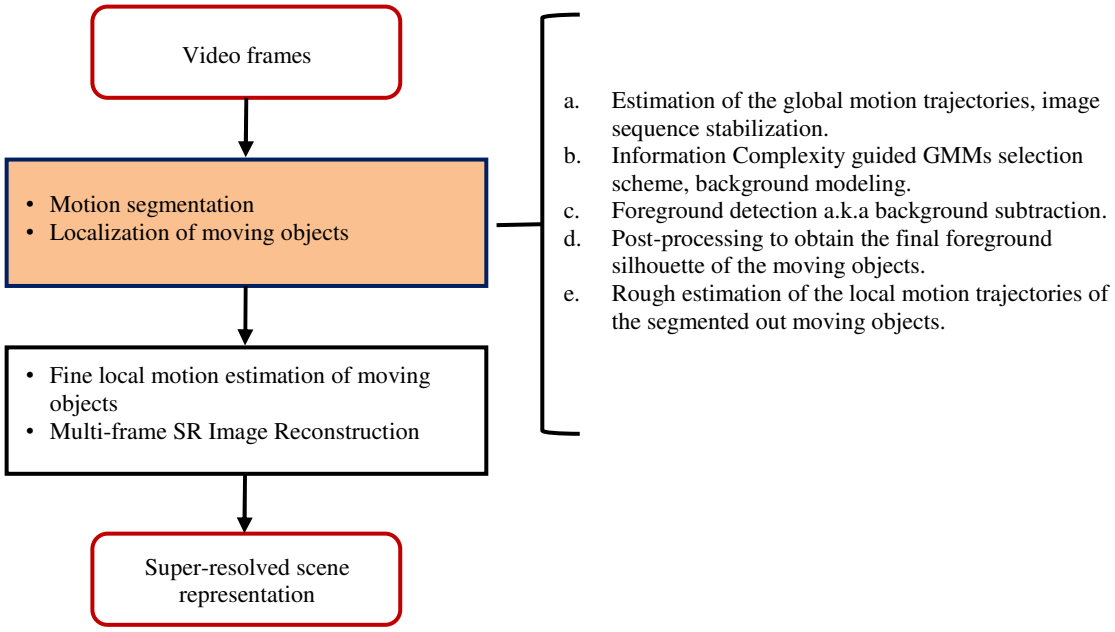


Figure 1.5: Block diagram of "Motion Segmentation aided Super Resolution (SR) Image Reconstruction".

Block of the framework which we contributed to.

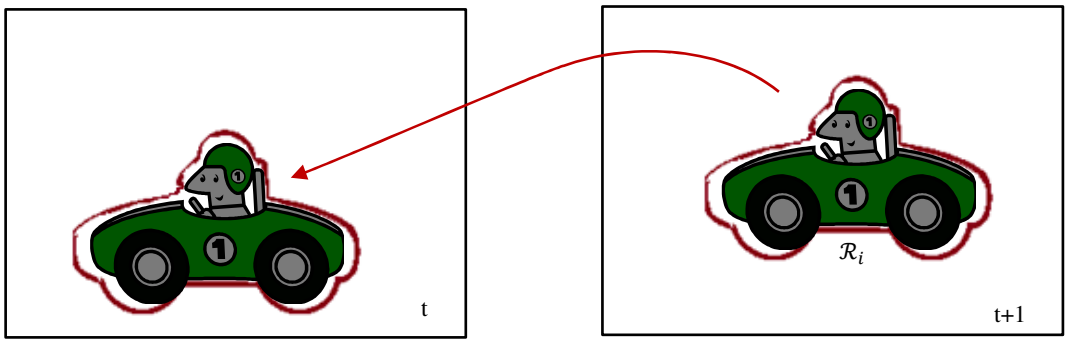


Figure 1.6: An illustration why motion segmentation is useful, region based alignment model.

Figure 1.7: The first goal, Extract out the regions which are parts of the moving objects and super-resolve it, (Source www.simplehelp.net).

For the Image/Video Reconstruction step; we use segmented out, localized moving regions in a two-step Traditional SR scheme. A refinement process of sub-pixel level fine local motion estimation of moving objects comes before the multi-frame Image Reconstruction. Ordinary Interpolation, Papoulis-Gerchberg [Papoulis77], Iterated Back Projection [Keren88], Robust Super Resolution [Zomet01], Projection onto Convex Sets (POCS) [Patti97], Structure Adaptive Normalized Convolution [Pham06] are the most commonly used algorithms for Image Reconstruction in the literature. Additional to these we also employ a wise-interpolation method Kriging [Krige51] which is a geostatistical tool and uses spatial characteristics of the data in a powerful manner. Geostatistics offers a way of describing the spatial continuity of natural phenomena (image acquisition) and provides adaptations of classical regression techniques to take advantage of this continuity. Kriging defines a stochastic process model, under which interpolation is done. We employ Kriging at the last stage as SR Image Reconstruction method to increase the resolution of the imaging system.

1.3 Contributions

Our goal is better understanding of the scene in monitoring and surveillance applications, in which an image sequence is provided to the user. From our point of view background has no importance; on the contrary, extraction of regions of interest out of a sequence of image is critically important. Stationary background information related to ordinary buildings, walls, roads, pavements, vegetation, sky, etc. are all excluded in our framework (Figure 1.7). Leaving out the background directly brings data redundancy to the framework which is the main reason we employ motion segmentation to aid SR Image Reconstruction. Accordingly, our research contributions are listed as follows.

Information Complexity guided Statistical Background Modeling: A new technique, Information Complexity guided Statistical Background Modeling is introduced in this study. Thus, we successfully utilized optimally computed GMMs, also known as Mixture of Gaussians (MoG) models, in background modeling to extract moving objects through background subtraction. Main contribution is shown in Figure 1.3.

Image Reconstruction of moving regions in non-stationary imaging systems: In addition, we developed a new framework of SR Image/Video Reconstruction of the moving objects, in which we are having high level of displacements of the moving objects resulting from not only the local motion of the objects but the global motion of non-stationary imaging system. In this framework, contrary to the traditional SR approaches we employed several steps to overcome the problems arising from high-level of misalignments. These steps are; Suppression of the global motion trajectories imposed on the image sequence, motion segmentation to extract moving objects, localization of moving objects through suppression of the local motion trajectories, super-resolving accumulated information coming from multiple LR frames to reconstruct an HR representation of the moving objects This framework proved to be superior to the state of algorithms which put significant effort for moving objects.

1.4 Document Organization

Following this introductory text, the remainder of this dissertation is organized as follows:

Chapter 2 gives a literature review of the topics most relevant to our research. Namely: Motion trajectories estimation, GMMs based Background Modeling, SR Image/Video Reconstruction.
Chapter 3 describes our efforts on global and local motion estimations.
Chapter 4 is the core theory chapter that develops the Information Complexity guided GMMs for Statistical Background Modeling. It argues experimentally the effectiveness of the scheme.
Chapter 5 presents the implementation of our framework SR Image/Video Reconstruction for the moving regions, segmented out from the video frames.
Chapter 6 contains a short summary of the dissertation's seminal points, a discussion with concluding remarks, and opportunities for future research.

2 Related Work

While the main thrust of this research is Information Complexity guided Gaussian Mixture Models (GMMs) for Statistical Background Modeling, other topics were also addressed in the process of developing our framework Motion Segmentation aided Super Resolution (SR) Image Reconstruction. These include the motion trajectories estimation towards global and local scene change detection, and image reconstruction to have high resolution (HR) representations of the extracted regions, which is the ultimate goal of the framework. In this chapter we present the relevant literature situating our work within the state of the art.

2.1 Motion Trajectories Estimation

Motion estimation in image sequences is an enormously big field and there is little point in attempting to present every prominent approach. We want to state what is necessary for us. Many strategies have been proposed and implemented for the image registration or motion trajectories estimation based on either the geometrical features (point-like anatomic features or surfaces) or intensity similarity measures. Knowledge of a 1-to-1 relationship between the grey value images is used in intensity similarity measures. Representation of the images for different kind of sensors invariant from brightness and contrast is usually not possible. Some examples of invariant image representations are edge maps, oriented edge vector fields, contour features, and feature points.

Reddy and Chatterji in [Reddy96] loosely divides registration methods into the following cases; algorithms that uses pixels values (e.g. correlation methods), algorithms that uses FT based methods, algorithms that uses distance transforms, algorithms that use low level features such as edges and corners (e.g. feature based methods), and algorithms that uses high level features such as identified parts of objects or relations between features (e.g. graph-theoretic methods). Irani and Anandan in [Irani00] name the first group as a as “feature-based methods” and the latter group as “direct methods”. Szeliski [Szeliski06] names the “direct methods” also as “pixel-based methods”. Feature-based methods minimize an error function based on the distances between sparse sets of feature points, and then recover and analyze their correspondences in order to determine motion and shape. Intensive methods based on ground control points or manually registered tie points have long been used to align images globally, and these are replaced with consistent solutions that can simultaneously solve the problem considering all the information we can gather from the image data. On the other hand, error measure related to the image information is collected from all the pixels of the images in direct methods. This information can be in different forms such as brightness gradients, and temporal changes as discussed above. Similarly instead of applying a refinement and regression utilizing brightness information, correlated images can be used in a correlation based direct method in [Irani98]. It is also possible to re-process the image before comparing their values, by using band-pass filtered images, or using local transformations such as histograms or rank transforms or mutual information obtained utilizing the regions can be maximized. Direct methods deals with minimizing pixel to pixel dissimilarities. Since direct methods confidence weighted local constraints from every pixel in the image to estimate a few motion parameters, these parameters are usually estimated to high precision, as a result the displacement

vector of each pixel by the motion is precise up to a fraction of a pixel. This has led a number of practical situations, one of which is super resolution image and video reconstruction, and sub-pixel alignment of the local structures or the entire images is a requirement. In this study we utilize global and local motion trajectories estimation scheme to accumulate all the moving region related information from LR images and present it as an input to the SR Image Reconstruction stage.

2.2 Gaussian Mixture Models based Background Modeling

The background subtraction techniques are common approaches for extracting foreground objects from image sequences through suppressing the information resulting from background. Background subtraction is just a small step following background modeling; therefore these two terms are loosely referring to the same practice. Background modeling is required to model the background and then detect the significant object regions in the scene in many imaging systems such as video surveillance, teleconferencing, video editing, and human-computer interfaces. The simplest way to model the background is just to take the frame which does not contain any moving object. However, in some environments, such as in aerial imaging one cannot keep recording for a long time to have the background with no moving objects. Moreover, the background can always be changed under critical situations like changes in illumination, objects entering to or leaving from the scene. To deal with the problems about adaptation to such circumstances, many background modeling methods have been developed and the most recent surveys can be found in [Piccardi04, Cheung05, and Elhabian08]. These background modeling methods are classified in the following categories by Bouwmans *et al.* in [Bouwmans10]: Basic Background Modeling, Statistical Background Modeling, Fuzzy Background Modeling, Neural Networks based Background Modeling, Background Modeling by Robust Principal Component Analysis (PCA), and Background Estimation. We will give details about intuitive Basic Background Modeling as well as Statistical Background Modeling approaches in Chapter 4. Background estimation (using e.g. Wiener filtering, Chebytchev filtering, Kalman filtering [Ridder95]), Fuzzy Background Modeling [Zhang06, El-Baf08], Neural Networks based Background Modeling, Background Modeling by Robust PCA are out of the scope of this study; they are just mentioned here to present a formal categorization of background modeling methods. Statistical Background Modeling or pixel modeling as named in [Friedman97], which considers a single pixel and the distribution of its values over time and extracts different states of it such as background, moving objects, shadows, etc., is the set point of the studies given here as the literature. A collection of statistical concepts for modeling the underlying data structure is sought in Statistical Background Modeling. Statistical variables are used to classify the pixels. Statistical model based background decomposition proves to be a useful tool for multivariate data [Stauffer99]. Therefore, in this study we carry our discussions on pixel-wise Statistical Background Modeling. All these categories are used in background subtraction which is directly related to the studies some of which are,

- Background initialization for complex dynamic scene analysis,
- Foreground detection,
- Choice of dominant pixel values for a given image sequence,
- Motion based segmentation,
- SR Image/Video Reconstruction of foreground objects, and
- Object tracking in dynamic scenes.

Yet, it is highly important; moving object detection schemes in complex environments is still not completely set [Elhabian08]. Also in the literature there are studies related to online models that focus on fast processing schemes during background modeling [Zivkovic04]. As reported in Elgammal et al. [Elgammal01] and Harville et al. [Harville01], there are several situations that must be taken care of by an efficient background subtraction algorithm to correctly extract moving objects. Relocation of background objects, non-stationary background objects (e.g. flags), and image changes due to camera motion which is

common in outdoor applications (e.g. wind load, bridge vibrations) should be considered. A background subtraction should be adaptive to illumination changes such as gradual changes (e.g. time of day), sudden changes (e.g. light switch), and global or local changes (e.g. shadows and inter-reflections). The situations related to the moving objects' characteristics should also be considered. When a foreground object might have similar characteristics as the background (e.g. the same texture as in camouflage), it become difficult to distinguish the objects from the background. It is not always the case that we have continuously moving objects. A foreground object can be motionless (e.g. sleeping person) or firstly moving then becoming motionless for the higher portion its existence in the scene (e.g. parked cars). In these situations separating it from a background is not achievable. In some adaptive Background Modeling studies a common problem faced in the background initialization phase is the existence of foreground objects in the training period, which occlude the actual background. On the other hand often it is impossible to clear an area to get a clear view of the background; hence this puts serious limitations on system to be used in highly dynamic scenes (e.g. high traffic areas). Some of these problems can be handled by very computationally expensive methods, but in many applications a short processing time is required.

Statistical Background Modeling algorithms have been developed to overcome this problem by modeling and updating the background statistics pixel-wise. They can be classified into two categories: parametric and non-parametric approaches [Kim07]. From now on we will use the term background modeling interchangeably with Statistical Background Modeling.

The non-parametric approaches estimate density functions directly from sample data. Elgammal used Kernel Density Estimators (KDE) to adapt quickly to changes in the background [Elgammal00]. They aimed to be able to accurately model the background process non-parametrically, so the model should adapt very quickly to changes in the background process, and detect targets with high sensitivity. Several advanced approaches using KDE were proposed. KDE based approaches are reported consume a lot of memory to update recent background statistics in [Kim05]. A codebook algorithm to construct a background model from long image sequences was proposed in that study. The researchers in [Stenger01] use Hidden Markov Models (HMMs) to switch states of the background with observations.

The parametric approaches set a parametric form of the background distribution (e.g. Gaussian distributions with μ, Σ parameters) in advance and estimate the parameters of the model. Pearson [Pearson94] is the first author to model a dataset consists of two populations with a GMM with two Gaussian distribution in 1894. For background modeling purposes earlier methods used single Gaussian distribution to model the probability distribution of the pixel intensity [Wren97]. Recently, GMMs is the most used approach in background modeling [Friedman97, Stauffer99] and has been extended in many studies.

2.2.1 Background Modeling using Gaussian Mixture Models

The Gaussian distributions are the most widely used tools for background modeling to detect moving objects from the image sequences. If non-stationary camera system is the case image sequences are globally compensated preceding background modeling. The probability of observing the current pixel value regarding the mixture of Gaussian densities is:

$$p(X_t) = \sum_{k=1}^{K} w_{t,k}\eta(X_t, \mu_{t,k}, \Sigma_{t,k})$$
$$\eta\left(X_t, \mu_{t,k}, \Sigma_{t,k}\right) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_{t,k}\right|^{1/2}} e^{-\frac{1}{2}(X_t-\mu_{t,k})^T \Sigma_{t,k}{}^{-1} (X_t-\mu_{t,k})} \quad (2.1)$$

where X_t is the current pixel value of the pixel history $(X_1 \ldots X_t)$, K is the number of the distributions, $w_{t,k}$ is an estimate of the weight, the portion of the pixel history represented by the k^{th} Gaussian

component in the mixture at time instant t, $\mu_{t,k}$ is the mean value of the k^{th} Gaussian component in the mixture at time instant t, $\Sigma_{t,k}$ is the covariance matrix of the k^{th} Gaussian component in the mixture at time instant t, $\eta(\cdot)$ is the Gaussian probability density function, p is the dimension of the each observation X_t. This model possesses the idea of updating the model at some time instances. The original idea was proposed by Stauffer and Grimson [Stauffer99].

The work of Friedman and Russel [Friedman97] can be regarded as the first example of what we call background modeling and the most similar preceding example to Stauffer and Grimson's work. They put their efforts on background subtraction for a road-watch traffic surveillance project. They proposed GMMs based classification for each pixel using an unsupervised technique, an efficient incremental version of Expectation Maximization (EM) to overcome the instabilities of standard time-averaging approaches a.k.a Basic Background Modeling. The mixture of three Gaussian components corresponding to road, vehicle and shadows are initialized using an EM algorithm. Their assumption is in the case of traffic surveillance, distribution of single pixel's values can be considered as the weighted sum of three distributions as p_r for road, p_s for shadow p_v for vehicle;

$$p(x_t) = w_r p_r(x_t) + w_s p_s(x_t) + w_v p_v(x_t) \tag{2.2}$$

They used a very heuristic approach when labeling these components. The darkest component is chosen as the shadows of the vehicles. For the remaining two components, the component with smaller variance is labeled as road and the other one is labeled as vehicle. Meaning of their approach lies in pixel modeling and a wise EM framework to train GMM. The behavior of their system for different types of pixels, which do not show characteristics of these three classes, is not clear. Naturally, a single pixel can have values resulting from other sources such as repetitive motions, reflectance, and daylight changes etc.

Stauffer and Grimson [Stauffer99] simply modeled the values of a particular pixel, namely the pixel history as a GMM. Based on the weight and the variance of each Gaussian distribution of the mixture, Gaussians corresponding to the background intensity values are determined and a background image is composed. Pixel values that do not match the background distributions are considered as foreground until there is a Gaussian that accepts them with sufficient, consistent evidence supporting it. Overall idea is separating the background from the foreground objects. Under conditions like lighting changes, repetitive motions of scene elements, tracking through cluttered regions, slow-moving objects, and introducing or removing objects from the scene they reported their method to be robust and efficient in background modeling task. They worked on an online adaptive background modeling, foreground object detection and classification system while monitoring the outdoor scenes for 16 months. For the sake of running an online fast responding background modeling system important parametric, temporal and spatial constraints were not dealt with. This method will be given in detail in Chapter 5 and the disadvantages due to assumptions they put on their system's behavior will be mentioned. The assumptions of that work, which are extended by many others later, are;

- The number of clusters K is blindly assigned,
- Covariance matrix for 3D (pixel values from RGB space) Gaussian components has the form $\sigma_k^2 \mathrm{I}$, where σ_k^2 are variances for each channel, and I is the identity matrix. Namely for any component the channels are assumed to be independent from each other. But these color components are surely dependent and so the simplification made there for the covariance matrix is not right.
- Online K-means approximation instead of an EM based -more robust- component update scheme is used,
- This algorithm does not distinguish shadows and objects. The normalized color module fails when the input signal has no color and its discrimination power is poor in dark and saturated areas.
- One other disadvantage, which is a general drawback for background modeling methods using a pixel-wise aspect, is that it prevents to handle some critical situations which can be only detected spatially and temporally.

– Furthermore, some critical situations need pre-processing or post-processing (e.g. camera motion).

In short, the original pixel-wise GMMs is designed well for time of the day, multi-modal background situations; medium for introduced objects, sleeping foreground objects, and it is not suitable for problems arising from camera motion, shadows, and camouflage. Such critical situations, as well as the real time constraints are investigated by many others and there is a good list of these methods in [Bouwmans10].

We want to continue our discussion on the augmentations made from the viewpoint of classic background modeling using GMMs. Intrinsic model improvements' concern is directly the initialization and the maintenance of the parameters. Additionally the robustness can be increased by adding the knowledge of external temporal and spatial process named altogether as extrinsic model improvements such as Markov Random Fields (MRF) to enforce temporal contiguity [Kumar00], spatial contiguity [Schindler06], Hierarchical approaches to combine pixel based and block based approaches [Chen07] two background models: one for color feature and one for the gradient feature [Javed02], Graph-cuts for shadow elimination [Sun06], strategies on employing complementary information coming from multi-modal image acquisition systems, such as combined IR and RGB features [Nadimi04].

We want to open a discussion on the color models proceeding to the intrinsic model improvements. Many studies (not in especially background modeling area) argue on that color is better than luminance for identifying objects in low-contrast areas and suppressing shadow cast by moving objects [Cheung05]. Generally the RGB space is used without modifying the data since RGB values are automatically provided by most frame grabbers. It is reported to be not well behaved in the context of color perception. The distance computed between two colors in RGB space does not reflect their similarity of informational perception [Elhabian08]. Wren et al. [Wren97] uses the YUV color space, and separates intensity (Y) and chromaticity components (U, V) in the pixel measurement. Similarly, the HSV model separates the intensity (V) from the chromatic components (H, S). However, the chromaticity representation based on linear combinations of R, G and B channels, is not as intuitive as the radial HS subspace representation [Francois99]. Elgammal et al. [Elgammal00] use the chromaticity coordinates as $r = R/s, g = G/s$ and $b = B/s$ where $s = (R + G + B)$ and $r + g + b = 1$. They claim this makes the background modeling advantageously insensitive to small changes in illumination that arise due to shadows namely it works as shadow suppression process.

2.2.2 Intrinsic Gaussian Mixture Model Improvements

We will continue our discussion based on intrinsic model improvements. Essentially the pixel modes describe the probability distribution of the appearance of the pixel conditioned on its type, where the type is the hidden variable [Friedman97]. In many studies pixel appearance is modeled as Mixture of Gaussians (MoG). However modeling the background does not always implies the assumption that distributions related to the background and foreground objects are Gaussians. Kim et al. [Kim07] show that the distribution of an indoor scene using Laplace model is more appropriate than using a Gaussian one. They used excess kurtosis given as:

$$g_2 = \frac{m_4}{\sigma^4} - 3 = \frac{n\sum_{i=1}^{n}(x_i - \mu)^4}{\left(\sum_{i=1}^{n}(x_i - \mu)^2\right)^2} - 3 \tag{2.3}$$

to measure whether the data are peaked or flat relative to a normal distribution. In the case of stable scenes such as indoor ones, variations of pixels are smaller than those in outdoor scenes due to less light dispersion and illumination change, and fewer of those small motions that tend to occur frequently in nature. Their suggestion for background modeling for indoor data is using Laplace distribution from generalized Gaussian family distributions instead of just Gaussians:

$$p(x;\rho) = \frac{\rho\gamma}{2\Gamma(\rho^{-1})}\exp\left(-\gamma^{\rho}\,|x-\mu|^{\rho}\right)$$
$$\gamma = \frac{1}{\sigma}\left(\frac{\Gamma(3\rho^{-1})}{\Gamma(\rho^{-1})}\right)^{1/2} \tag{2.4}$$

where $\Gamma(\cdot)$ is a gamma function. For $\rho = 2$, $p(x;\rho)$ becomes a Gaussian distribution, whereas for $\rho = 1$ it is a Laplace distribution. In another way, Wang et al. [Wang05] reports that when the intensity varies abruptly the intensity does not follow the Gaussian distributions, like in the case of flickering trees of outdoor scene. Therefore, if the goal is not just to compress the dataset but also to make inferences about its distribution structure, it is essential to analyze whether the dataset exhibits a clustering tendency. Furthermore it can be difficult to tell from data whether a physical or other observed process is random or chaotic. Trying to cluster a chaotic data is not possible. Yet, behavior of image acquisition process in a time period can be regarded as stochastic rather than chaotic. From time to time we do not expect exponential changes for any pixel.

For the initialization, the GMMs asks for the number of Gaussians, K. It is mostly fixed and the same for all pixels. However, this assumption is not optimal because the multi-modality is variable spatially and temporally. In pixel-wise background modeling form of the covariance matrix, as well as the number of the components to represent the probability distribution of each pixel may vary. For the initialization of the mean, the variance and the weight, a series of training frames absent of moving objects is needed but in some environment, it is not possible to obtain such frames. For the adaptive parameter update maintenance phase, Greiffenhagen et al. [Greiffenhagen01] characterizes it statistical behavior making different parameters' initialization. To characterize the statistical behavior of background adaptation module, numerous experiments on real data as well as on simulated data were conducted. Random samples from a mixture distribution of components with model parameters w_i, μ_i, Σ_i with $i \in \{1,2,3\}$ were generated. They observed how the model parameters typically evolve over 10,000 time intervals. The experiment shows that only the means are estimated and tracked correctly. Up to their evaluation, variance and the weights are unstable and unreliable. They also augmented the study of Stauffer and Grimson to handle shadow information and by proper statistical fusion of the two modeling schemes: modeling the gradual changes in background due to the illumination spectrum and non-linear dynamics, and modeling the changes change due to sudden camera gain/shadow. By using classical algorithm first and feeding its internal state to the normalized color change detection algorithm, they reported gaining the advantages of both.

The number of clusters K is intentionally, blindly assigned to run an online background modeling system, and fixed to 3 or 5 for each pixel in [Stauffer99], therefore this number is not optimal. [Friedman97] uses K as 3 and label them as vehicle, background shadow and the state of the pixels resulting from repetitive motions, reflectance, daylight changes etc. are all included one of this three. Many others just select the K number as 3 or 5 [Pavlidis03, Amintosi07]. To solve this problem, [Zivkovic04] proposed an online algorithm that estimates the parameters of the GMMs and simultaneously selects the number of Gaussians using the Dirichlet prior. The outcome is that not only the model parameters of the Gaussians but also the number of clusters K are dynamically adapted to the multi-modality of each pixel. By choosing the optimal number of components for each pixel in an on-line procedure, the algorithm reported to be automatically fully adapting to the scene. Carminati et al. [Carminati05] estimate the optimal number of K Gaussians for each pixel in a training set using an ISODATA algorithm. This method is less adaptive than the others because K is not updated after the training period. For the same objective, Cheng et al. [Cheng06] develops a stochastic approximation procedure which is used to estimate the parameters of GMM recursively. The number of Gaussians obtained reported to be asymptotically optimal. Shimada et al. [Shimada06] proposed an approach consisting steps to dynamically control of the number of Gaussians. This approach changes the number of Gaussians for each pixel automatically. This number increases when the variance for the current pixel history is high, to control a larger space for classification. On the other hand, when pixel values are

constant for a while namely the variance is very low; some Gaussians are eliminated or integrated to the current ones. This process helps reduce computational time. Tan et al. [Tan06] proposed a background modeling called Adaptive-K Gaussian Mixture Model (AKGMM). It was about traffic Video Segmentation to classify pixels in the current frame as road background or moving vehicles, and casting shadows if observed. Their framework comprises a modified online EM procedure to construct an adaptive GMM in which the number K can adaptively reflect the complexity of pattern at the pixel. A simple shadow detection algorithm called Normalized Cross-Correlation algorithm (NCC), proposed by Julio et al. [Julio05] to refine the segmentation if dynamic casting shadow exists, was utilized. They reported their method to have the capability of detecting shadows using NCC.

2.3 Super Resolution Image Reconstruction

As stated in the Chapter 1, SR methods stem back to single image restoration problem long before modern multi-frame SR methods became prominent. SR algorithms have been proposed by many authors to reconstruct reliable representation of the scene for further recognition and understanding purposes. Good overviews of the algorithms are given in [Katsaggelos07, Park03]. To present the methods in a consistent fashion several categorizations which are based on different aspects such as models, reconstruction strategies, domains, and etc., will be presented here in this study.

To begin with, the authors in [Kim10, Yu08, Glasner09] states two SR categories based on whether a training stage is employed in SR or not, as learning based SR methods and reconstruction based SR methods. The underlying idea of the first group is to learn a map from input low resolution (LR) images to HR images based on example pairs of input and output images [Kim10]. HR information is assumed to be split up among multiple LR images, implicitly found there in aliased form. In learning based SR, this missing HR information is assumed to be available in the LR database patches, and learned from LR/HR pairs of examples in the database and then applied to a new LR image to recover its most likely HR version. Although these methods have already shown impressive performance, it is well known in computer vision community that regression based estimations suffer from over-fitting when the target function is highly complex or the data is high-dimensional, which is the case for SR. Accordingly, it is reasonable to expect that nearest neighborhood based methods can be improved by adopting learning algorithms with regularization capability to avoid over-fitting. These approaches relies heavily on having large database of HR images under varying pose and illumination conditions, and not feasible for all application scenarios. Learning based SR has been shown to exceed the limits of edge learning models. However, unlike classical SR, the HR details reconstructed or 'hallucinated' by learning based SR are not guaranteed to provide the true (unknown) HR details [Glasner09]. Reconstruction based methods utilizes additional observation data along with spatio-temporal observation constraints mainly obtained from sub-pixel motion compensation process. Glasner et al. [Glasner09] name their work as single frame SR method. Their approach is based mainly on recurring image patches, both within the same scale, as well as across different scales. Instead of using multiple frames taken at different time instants, they use a single frame and use multiple image patches all observed at the same time. Unlike single image restoration, such a method still tries to use abundant information from multiple patches and produces each patch highly resolved at different scales. For the remaining of the section reconstruction based SR methods will be referred with the term SR methods.

Many SR studies reconstruct only a single HR frame from various LR frames. The process of can be applied to reconstruct the image sequence but it does not take the advantage of any previously estimated HR frames. In [Zibetti05, Borman99] the authors classify the SR methods as; sequential SR methods which estimate the HR frames at one time, using many LR frames and other HR frames previously estimated, and simultaneous algorithms which estimate the entire sequence where all HR frames are restored, in one process. Zibetti and Mayer [Zibetti05] proposed a simultaneous algorithm to estimate the entire image

sequence based on maximum a posteriori (MAP) estimation in contrast to the other multi-frame SR algorithms. They preferred not to include the motion in the observation model. They used the motion as a priori information in order to achieve smoothness along the motion trajectory.

In [Juan09, Vandawalle06, Keren88] SR methods are divided into two broad categories as the frequency and the spatial domain methods and they put all efforts under either of these two. Motion estimation and image reconstruction are two required stages in many SR techniques. Utilizing domain based techniques at one of these stages forces us to classify the method as either spatial domain method or frequency domain method. Great majority of SR methods fall under the spatial domain methods. In this broad class, the observation model is formulated, and reconstruction is employed in the spatial domain. The linear spatial domain observation model can accommodate global and non-global motions, and can compensate the effects of spatially varying phenomena [Borman98]. The spatial domain reconstruction allows natural inclusion of (possibly nonlinear) spatial domain a-priori constraints (e.g. Markov random fields or convex sets) which result in bandwidth extrapolation in reconstruction. Bayesian theory based SR methods provide a powerful theoretical base for the inclusion of a-priori constraints necessary for the solution of ill-posedness in SR. In a random process, a system's subsequent state is determined both by the process's predictable actions and by random elements. A simultaneous multi-frame super resolution procedure, utilizing spatio-temporal smoothness constraints and motion estimator confidence parameters was proposed by Borman and Stevenson; degraded observations are formulated as a statistical inference problem and a MAP approach was utilized [Borman99]. Schultz and Stevenson [Schultz94] are the pioneers to formulate MAP approach to estimate the HR resolution image using single LR frame. They proposed the idea of inclusion of the additive Gaussian noise. For the noise free case of $\mathrm{g} = \mathrm{Df}$, where g is $MN \times 1$ lexicographically ordered vector that contains pixel values from the LR image and f is the $q^2MN \times 1$ vector containing pixel values from the HR image −with $q > 1$ the relative sensor size of the HR image acquisition system with respect to that of LR image acquisition system−, and the decimation system model D is the $MN \times q^2MN$ size decimation matrix. The image acquisition system is stable $p(\mathrm{g}|\mathrm{f}) = 1$ for $\mathrm{g} = \mathrm{Df}$. For same HR image same LR will be obtained from the imaging system with no randomness. For the case involving additive Gaussian noise, constraints towards approximating the HR image can be obtained using noise vector n which is a random process in an imaging system. To model the a-priori function MRF is assumed with Gibbs density function in [Schultz94]. An extension of simple frame method to multiple MAP estimator was developed in [Schultz96] by the same authors.

A major class of SR methods utilizes a frequency domain formulation to solve the SR problem. Frequency domain SR methods provide the advantages of theoretical simplicity, low computational complexity and they are highly amenable to parallel implementation due to decoupling of the frequency domain equations and exhibit an intuitive de-aliasing SR mechanism. Each LR image contributes independent structures which governs the inter-frame motion in the frequency domain was the idea implemented first in SR history, by Tsai et al. [Tsai84]. Their observation model was based on the shift property of FT, observed Landsat images are modeled as under-sampled versions of unchanging scene undergoing simple global translation. Their work disregarded both the blur and the noise in the imaging process. A frequency domain technique was proposed by Vandawalle to precisely register a set of aliased images based on low frequency, aliasing free part. The resolving power can also be increased by bringing in high frequency information based on the image model or by removing the aliasing ambiguity [Vandawalle06], a high resolution image then reconstructed using simple cubic interpolation. One main advantage of employing the frequency domain methods is indicated as, if the one image is the shifted version of another than a phase shift can be seen in the frequency domain. Using a log polar transform of the magnitude of frequency spectra image rotation and scale can be converted into horizontal and vertical shifts. Marcel et al. [Marcel97] and Reddy and Chatterji [Reddy96] described such planar motion estimation algorithms. Horizontal and vertical shifts can be estimated separately from the rotation. Yang and Schonfield in [Yang09], investigate to improve performance analysis of Super Resolution (SR). They derived lower bounds on the resolution enhancement factor based on a frequency domain SR algorithm and discussed the extension of the performance bounds

to temporal Super Resolution methods and its implications on the image sequence. From this respect under kernel based SR methods we can include two more groups of related approaches; discrete cosine transform (DCT) based SR method was proposed by Rhee and Kang [Rhee99]. They reduced memory requirements and computational costs by using DCT instead of DFT.

A member of domain based methods such as spatial and frequency domain methods; wavelet based SR methods are employed many times recently [Wheeler07, Whillet03, Nguyen00]. The ideal algorithm for SR should be fast, and should add sharpness and details, both at edges and in regions without adding artifacts. In [Hsu04] a wavelet based super resolution study is divided mainly into three stages; image registration, wavelets based fusion and image deblurring. The wavelet based fusion is performed to overcome the need for retaining edges like details when going from LR images to HR images. Such techniques reduces blocking artifacts, highlights the edges and it is also able to restore high frequency details in an image. Wavelet analysis is employed in reconstruction step for denoising, and accurate and sparse representation of images consisting smooth regions. Discrete Wavelet Transform (DWT) performs poorly when the frames contain motion like blurs. High frequency coefficients across all frames are combined by a fusion scheme. Fusing the high frequency information reduces the erroneously enhanced noise seen after deconvolution. Original reference image (the LR image which the other LR images are aligned onto) is up-sampled using interpolation and DWT is applied. High frequency coefficients are replaced with the fused high frequency coefficients, which is the essence of the studies employing wavelet based super resolution [Wheeler07, Hsu04].

Sroubek et al. [Sroubek07] proposed a very sophisticated method, considering all the components of the observation model shown in Figure 2.1 formulated as:

$$g_l = DV_l W_{l,k} f_k + n_l. \tag{2.5}$$

where n_l denotes the acquisition and registration noise, $W_{l,k}$, is the warping matrix warps k^{th} HR image and creates $W_{l,k}f_k$, V_l represents the volatile blurring D is the down-sampling or decimation matrix, g_l is the l^{th} LR Image.

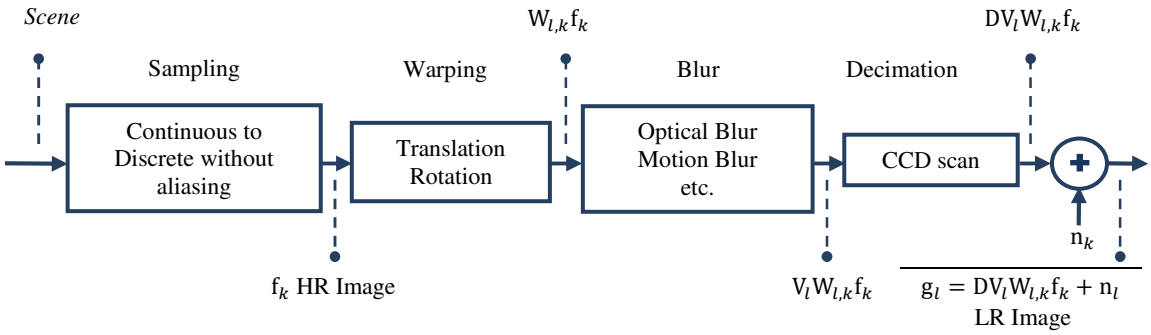


Figure 2.1: Temporally non-coincident warp-blur observation model.

3 Motion Trajectory Estimation

Motion estimation of the structures in image sequences is often the first step required for such diverse applications including moving object detection and tracking, robot navigation, vehicle control, environment mapping, motion based video coding and compression, as well as Super Resolution (SR) Image/Video Reconstruction. Since motion estimation is an enormously big field and researchers have produced great amount of related methods, there is little point in attempting to present every prominent approach.

Motion segmentation through background modeling can be applied directly to LR image sequences acquired with stationary camera systems. In the case of non-stationary camera system, the simple provision of moving image areas to the segmentation algorithm will not be sufficient. Instead, in order to separate objects from background, information about the camera motion is expected to be available with the help of initial global motion estimation of the frames. We prefer to put this chapter before the one on motion segmentation considering that our early practices are related to non-stationary camera system arisen challenges. Image acquisition platform's instabilities cannot be compensated, and we cannot go further to moving object segmentation without the aid of global motion estimation. Finding out relevant parts of the scene in the temporal domain through motion segmentation gives us a huge comfort in terms of finding the moving region correspondences in different images for further tasks of local motion estimation. The discussion in this chapter is carried out such that we attempt to clarify the fundamental character and challenges of the motion estimation problem, also its use in our framework. The following discussions on motion analysis are guided by the needs of motion segmentation as well as SR Image/Video Reconstruction part of this study.

3.1 3D Motion, Projected Motion and Optical Flow

In imaging systems, the 3D relative movement of both objects and camera is induced as 2D motion on the image plane via a suitable projection system. We want to estimate the 2D motion (in the forms of velocity or displacement) field from time varying images. However, what we perceive is the apparent motion (in the forms of optical flow and correspondence) field. The correspondence and optical flow fields are respectively displacement and velocity functions perceived from time varying image intensity pattern [Tekalp95]. The estimated motion is typically described using instantaneous velocity fields or correspondence fields (Figure 3.1). With the help of intensity variation information from the images, projected motion can be recovered apparently, not actually. The data is often degraded by noise and disturbances of different nature; motion estimation seeks for the correct 2D movements in the image plane [Ercole04].

A point $\mathrm{X}(t) = [X(t), Y(t), Z(t)]^T$ on a moving region in 3-D space can be projected onto the camera's focal plane at position $\mathrm{x}(t) = [x(t), y(t)]^T$. The projections of a 3D point on the image plane at a time instants t and $t + \Delta t$ $(\Delta t > 0)$, at positions $\mathrm{x}(t)$ and $\mathrm{x}(t + \Delta t)$ will correspond to each other, where ΔT is the time interval. Displacement or correspondence vector can be described as $\mathrm{d}_{t,t+\Delta t}(\mathrm{x}) = \mathrm{x}(t + \Delta t) - \mathrm{x}(t)$ on the image plane. In this case, image values can be predicted (from a past reference frame) using the

assumption of brightness constancy that $g_t(\mathrm{x}) = g_{t+\Delta t}\left(\mathrm{x} + \mathrm{d}_{t,t+\Delta t}(\mathrm{x})\right)$ where g_t and $g_{t+\Delta t}$ are two different dependent image functions at two time instants t and $t + \Delta t$. $\mathrm{v}_t(\mathrm{x}) = \mathrm{d}_{t,t+\Delta t}(\mathrm{x})/\Delta t$ will be the instantaneous velocity (optical flow). Optical flow vector is defined as the temporal intensity change rate of the image and it is equivalent to correspondence vector assuming the velocity remains the same during each time interval. In Figure 3.1, high order motion trajectory is illustrated with a curve. Motion trajectories, if approximated accurately can sometimes be used to reconstruct images between temporal sampling instants yet interpolation or extrapolation in the temporal domain is not in the focus of this study.

In an imaging system we can observe the spatio-temporal variation of the light intensity occurrence at the image plane. Interaction of the scene illumination with the objects in the scene, motion of the objects in the scene, and changes of camera's extrinsic parameters (position, orientation) or intrinsic parameters (focal length, focus setting, etc.) leads to spatio-temporal variations. Not all changes in the image intensity correspond to scene motion, nor does all scene motion result in image intensity variation. For example, changes in scene lighting result in image changes which do not correspond to any 3D motion. On the other hand, a uniformly illuminated disk having an axial rotation, which is a definite 3D motion, does not produce any observable change in image intensity at the image plane [Borman02]. The optical flow is zero at all points in the image. Despite these difficulties, using the time-varying intensity information, it is still possible to approximate the optical flow field. We assume illumination is uniform all across the surfaces; reflectance varies smoothly and has no great spatial discontinuities. Thus for convenience, the apparent motion estimation for a sequence of images, which we study here, can be directly identified with the movement of surfaces in the scene.

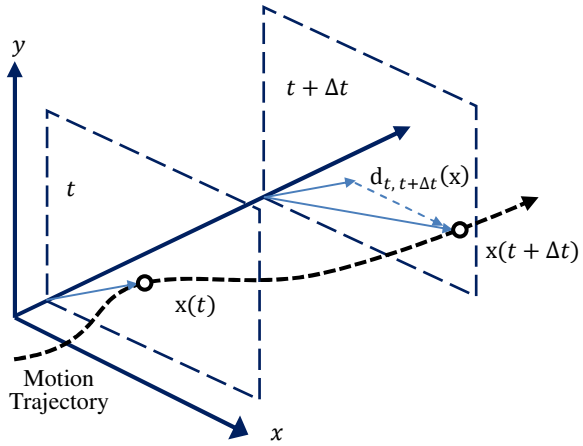


Figure 3.1: Motion trajectory and motion estimation using two frames, re-plotted from [Borman02].

3.2 Spatio-temporal Motion Analysis and Motion Segmentation

Motion estimation requires finding correspondence between image regions (in forms of displacement or velocity vectors) undergoing different levels of movement from one frame to the next. Potential problems such as blurring, varying image exposures, electronic noise induced at the imaging sensors should be taken care if possible during or before the motion estimation process. Many of the image sequence processing methods begin with the attempt of computing the optical flow maps which shows how the image regions are changing with time. Some researchers [Irani00, Szeliski06] group motion analysis/estimation attempts under two large categories. Approaches that use pixel-to-pixel matching are often called direct methods, as opposed to the feature-based methods. Pixel-based (direct) methods use all the pixels within the region of support and eliminates the necessity of salient structure (such as the edges, corners) extraction. In fact, distinct features are helpful yet they are not easily detectable. Some methods extract a sparse set of distinct features from each image separately, then recover and analyze their correspondences in order to approximate the motion. A widely accepted but not the most general categorization discriminates the different motion analysis/estimation techniques into three classes; gradient-based, correspondence-based, and frequency-based approaches [Kuhne02]. These approaches are used for computing merely the optical flow using different means.

Gradient-based methods —estimate motion fields by calculating spatial and temporal derivatives of image intensities. As mentioned before assumption of brightness constancy should be made when attempting to determine optical flow. Gradient-based methods demand small displacements because large displacements make the anticipated accurate numerical differentiation impractical. This assumption is implicitly embedded in a wide variety of motion estimation techniques even though the formulation of this constraint differs for technique to technique. Denoting the time varying image intensity function at location (x, y) at time instant t as $I(x, y, t)$ and the change of the intensity after a small movement as $I(x + \delta x, y + \delta y, t + \delta t)$ the optical flow equation can be stated as:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t + O(n^2)$$

$$I(x + \delta x, y + \delta y, t + \delta t) \cong I(x, y, t) \tag{3.1}$$

$$\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t \cong 0 \;\longrightarrow\; \frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t} \cong 0$$

$u = \delta x/\delta t$ and $v = \delta y/\delta t$ are the vertical and horizontal components of the optical flow vector $\mathrm{v} = [u, v]^T$ respectively, and the gradient vector $\nabla I = [\partial I/\partial x \;\; \partial I/\partial y\,]^T$ is composed of spatial derivatives of the image brightness. We can rewrite Equation 3.1 as:

$$\nabla I \cdot \mathrm{v} + \partial I/\partial t \;= 0 \quad \text{or} \;\; \langle \nabla I, \mathrm{v} \rangle + \partial I/\partial t \;= 0 \tag{3.2}$$

All the quantities in these equations are functions of image positions (x, y) hence every pixel provides one such equation. It is not possible to determine the local motion without any additional constraints [Irani00]. A more specific challenge statement is delayed to the coming discussions not to abrupt the flow of motion estimation categorization.

Correspondence-based motion analysis —identifies corresponding image structures in consecutive frames. Appropriate image structures can be listed as unprocessed image regions, image blocks of a any size, corners, edges, etc. Those structures can be matched in consecutive frames in different ways, e. g., by relying on a two-dimensional search within a window, graph theoretic methods, or relaxation labeling.

Fourier analysis — utilizes of periodicity of the patterns in an image. Using of Fourier Transform (FT), an image is projected onto complex exponential components which form an orthonormal basis, and the projections reveal the spatial frequency spectrum of the image. Frequency domain methods contribute to the solution of the problem with the light of the basic principles; the shifting property of FT, aliasing relationship between the Continuous Fourier Transform (CFT) and the Discrete Fourier Transform (DFT) [Borman98]. These properties lead to a methodology relating the aliased DFT coefficients of the observed images to the samples of CFT of the unknown scene. One main advantage of employing frequency domain methods is indicated as, if one image is the shifted version of another than a phase shift can be detected easily in the frequency domain. Using a log polar transform of the magnitude of frequency spectra image rotation and scale can be converted into horizontal and vertical shifts. Marcel et al. [Marcel97] and Reddy and Chatterji [Reddy96] described such planar motion estimation algorithms. Horizontal and vertical shifts can be estimated separately from the rotation.

How to Overcome Aperture Problem?

The solution set given in Equation 3.2 produces infinite number of u, v values. We can only be accurate about the projection of the optical flow onto the intensity gradient ∇I, such that:

$$\mathrm{proj}_{\nabla I}\mathrm{v} = \mathrm{v}_{\perp} = \frac{\langle \nabla I, \mathrm{v} \rangle}{\|\nabla I\|} = \mathrm{v} \cdot \frac{\nabla I}{\|\nabla I\|} = -\frac{\partial I / \partial t}{\|\nabla I\|}. \tag{3.3}$$

It is clear that given the intensity gradient and the temporal partial derivative only, the normal component of the optical flow $\mathrm{v}_{\perp}$ can be estimated as illustrated in Figure 3.2. This limitation is often referred as the aperture problem and it is under-constrained. All optical flow methods introduce additional conditions for estimating the actual flow. As, described above, the aperture problem might cause parts of the objects to be left out. Due to the areas of constant gray values within the moving objects, we normally do not receive dense motion vector fields. The identified regions of a moving object accordingly contain gaps and holes. Consequently, to extract out the complete locally moving regions from the video sequence, a grouping step is needed that integrates local information obtained from the motion detection algorithms. Ideally, such a grouping step should fulfill a number of properties: The final contour which separates objects from the background should reproduce the boundaries apparent in the image. Furthermore, missing parts should be approximated naturally, and the grouping step should be able to find several objects simultaneously. Active contour models, which are also known as deformable models, snakes (in 2D), or active surfaces (in 3D), are widely used in the problem domain of grouping local information.

Motion segmentation as a process itself is proved to be a great solution to detect the presence of motion discontinuities and to prevent false detection of motion at certain image regions. Given an image sequence from a fixed image acquisition system, separating all moving objects is the main essence of the Motion segmentation techniques. Motion segmentation involves with change detection to segment each frame into regions as well as motion estimation to find correct correspondences if the camera system is non-stationary.

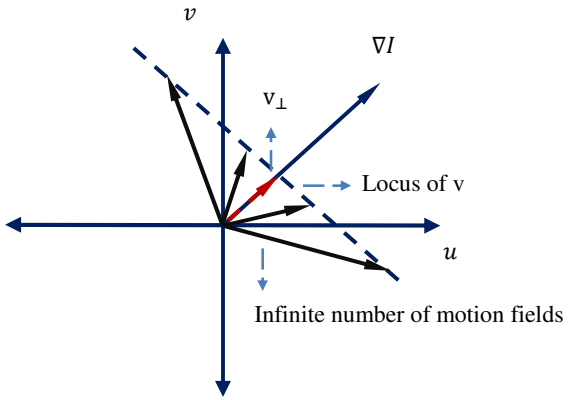


Figure 3.2: Aperture problem, under-constrained solution of optical flow.

3.3 Motion Field Representations

For motion estimation, we must first determine the appropriate mathematical model (motion models accompanies with a designated region of support) relating pixel coordinates in one image to pixel coordinates in another image of the sequence. Motion field representations may be divided into two broad categories [Borman02] as non-parametric and parametric motion models, each having distinct advantages and disadvantages. In non-parametric motion field models, a representation of the motion field is pursued on a finite set of points in the 2-D image plane. The primary advantage of this approach is that arbitrary motion fields may be represented. The motion field may be interpolated conveniently to produce values between sampling points. The main disadvantage of the non-parametric representation is that it requires the estimation of a large number of motion parameters. This makes non-parametric models poorly suitable in some image sequence applications. The other category is parametric motion models which represent the motion field over some regions of the image. Common 2D and 3D parametric motion models use 2 to 15 parameters which are summarized in Table 3.1 [Fitzgibbon03, Szeliski06]. 2D planar parametric motion models are illustrated in Figure 3.2. Once the parameters and the region of support of the model are determined, the model may be evaluated at any location x within the region, thus there is no need for interpolation. Parametric models have the advantage of requiring relatively fewer model parameters to describe large motion fields. Since the number of model parameters is small, this tends to yield more reliable estimates. Parametric models have disadvantages either; arbitrary motion fields cannot be represented using parametric models. Increasing the number of model parameters makes the model similar to the non-parametric models. Estimation of the region of support of the parametric model can be very difficult for general motion fields. Since the region of support of the non-parametric model is a point, this problem is not encountered. In SR field to estimate the 2D projected motion simple parametric models are used but the complexity of the interdependency of the frames leads us usually to non-parametric representations. Currently, there is no motion estimation approach that work reliably for all kinds of motions. For instance, generic techniques will fail for scenes with effects such as inter-reflections, specularities, and translucency [Khan06]. Motion estimation related to Lambertian objects the surfaces of which reflects the light the same in all directions cannot even easily handled in the case of exposure changes. Having dynamic range images through exposure changes during video acquisition is mostly omitted so as to achieve maximum application independence. An important term which goes hand in hand with motion estimation model is the region of support. Once the support of the motion estimation is described we can model the displacement of every pixel in that region of support. Various partitions of the

image plane into regions on which parametric models are models are applied will be discussed in this section. The partitions of the image plane $\mathcal{R}$ can be denoted as $\{\mathcal{R}_i\}_{i=1}^{N}$ such that $\bigcup_{i=1}^{N}\mathcal{R}_i = \mathcal{R}$, $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$. Global models use the partition $\mathcal{R}_1 = \mathcal{R}$ that the region of the support for global motion is the entire image plane. When describing camera motions, such as translation, rotation or zooming on a scene without moving objects, global models are the most useful. Block-based models are the most instinctive way of motion estimation; the partitions $\mathcal{R}_i$ are equal sized rectangular blocks, a parametric motion model applies for each block. Block-based models are attractive to a great number of applications; they are poorly suited to the task of accurately describing general motion fields. The fixed size blocks can be more functional using adaptive triangular meshes or hierarchical blocks. $\mathcal{R}_i$'s are triangles or blocks of various sizes [Borman02]. Normally there is no restriction of having irregular shape regions. In region-based motion models $\{\mathcal{R}_i\}_{i=1}^{N}$ may have on arbitrary shapes. Determining the arbitrary shaped $\mathcal{R}_i$'s is a difficult undertaking. Moving regions if available provide unavoidable cues to form region of support for region-based motion models. Finally one can say non-parametric motion models as the extreme case for of parametric motion models where the region of support is a single point $\{\mathcal{R}_i\}_{i=1}^{N}$ where N is the number of pixels.

Table 3.1: Parametric motion models commonly occurring in the literature.

n-D Transformation	Preserves	Degree of freedom, Number of model parameters
2D Translation	Orientation	2
2D Euclidean (Rigid)	Euclidean distances	3
2D Similarity	Angles	4
2D Affine	Parallelism	6
2D Projective	Straight Lines	8
3D Translation	Orientation	3
3D Euclidean	Euclidean distances	6
3D Similarity	Angles	7
3D Affine	Parallelism	12
3D Projective	Straight Lines	15

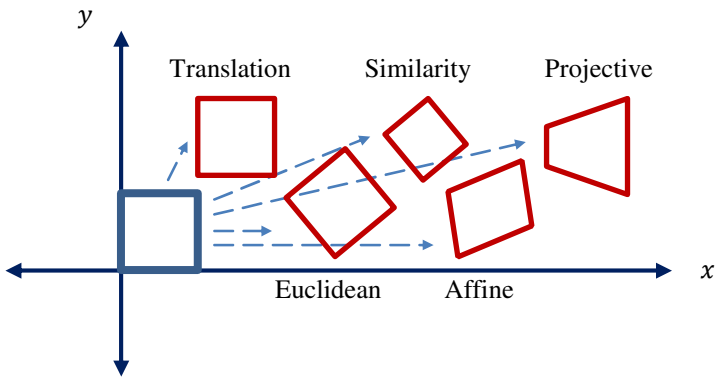


Figure 3.3: Basic 2D planar parametric transformations.

Different motion estimation models and structures of region of support can be utilized. In this study, we are dealing with both the camera instabilities and the movements of the objects observed on the dynamic scene. Contrary to the traditional SR approaches, we employed several steps to overcome the problems arising from high-level of misalignments. The following tasks are directly related to the base constructed by motion estimation efforts; suppression of the global motion trajectories imposed on the image sequence, motion segmentation to extract moving objects and bring the convenience of utilizing region based motion models, localization of moving objects through suppression of the local motion trajectories of the moving objects. Determination of the arbitrary shaped $\mathcal{R}_i$'s is a difficult undertaking and in this context motion segmentation helps us to employ region based alignment models.

3.4 Experimental Results

The segmentation of moving objects in images becomes harder when camera itself is moving or the platform, on which the camera is mounted, is moving. In this section of the dissertation, we are dealing with the methods of global motion estimation as a pre-processing to background modeling which will be discussed in detail in the following chapter. Our objective here is to distinguish moving regions from the background which has a global motion due to non-stationary behavior of the imaging system.

Construction of global motion estimation can be described in the following steps;

- We adopt block-based model, use equal sized rectangular blocks and put a parametric motion model for each block.
- For each block a block variance score is assigned to prevent unnecessary use of simple textured blocks such as clouds, big walls etc. Doing so the blocks giving a high variance score are favored.
- After deciding on the good blocks to use, first a translational model is sought to estimate misalignments related to each block. A displacement vector map is created.
- For this map the variance of both the horizontal and vertical displacements are calculated, if the variance of the displacements along two dimensions is below a threshold then translational model is accepted, otherwise for each block additional to the translation parameters a rotation parameter is sought.
- Obtaining the parameter maps for blocks, median values are calculated to represent the overall global motion observed on consecutive frames. Through such a process we avoid using the information coming from the moving objects. The changes arising from locally active moving objects are intentionally dismissed.

To use such a motion estimation scheme described above there is a certain assumption we put related to the images; moving objects are relatively small compared to the background. Also the motions are computed using consecutive frames, and a solid stabilization in which the first and the last frames are close to each other globally is not pursued. Bringing just the consecutive frames to the same coordinate system is the objective. An illustrative overview of the global motion estimation process is given in Figure 3.4.

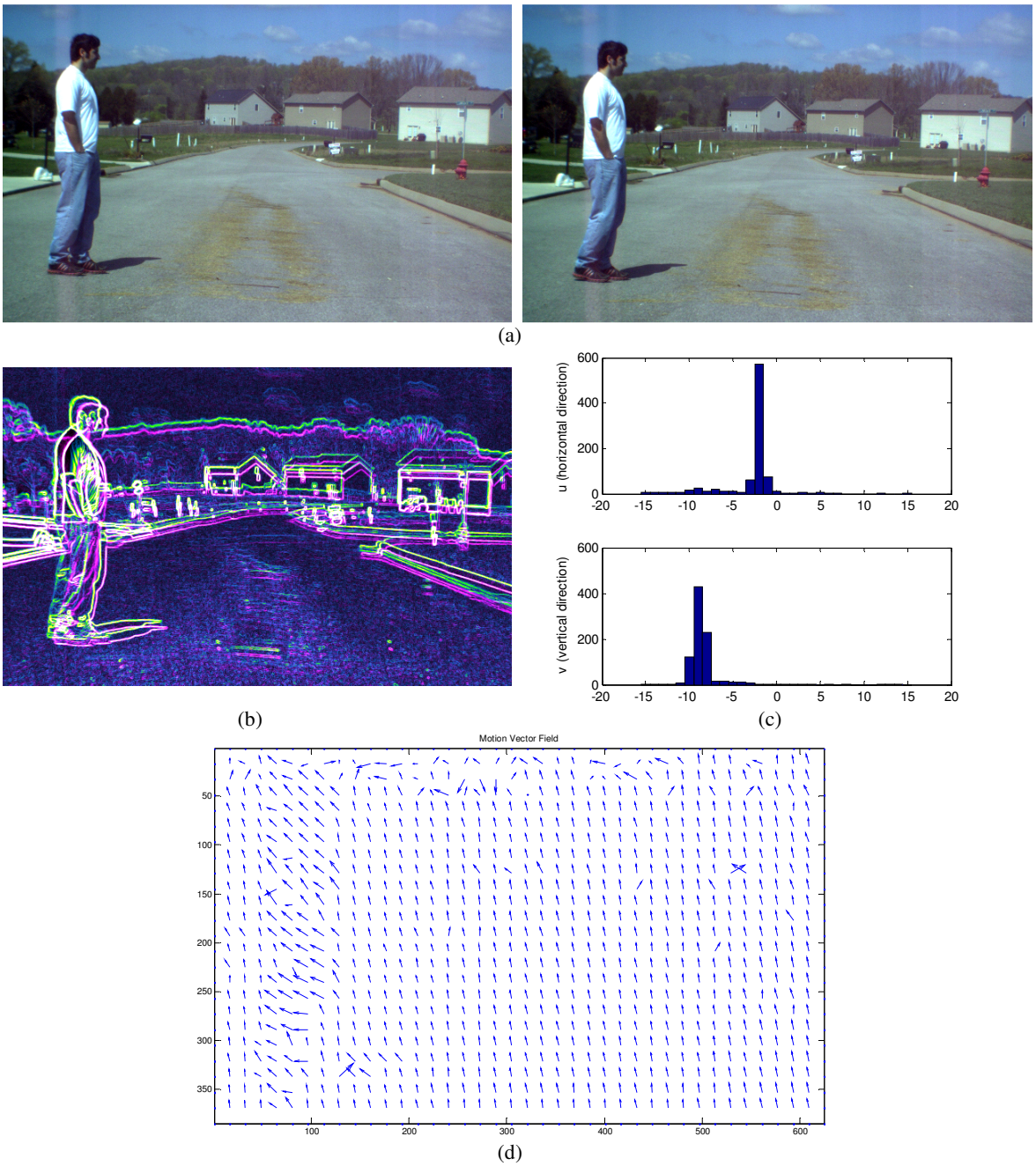


Figure 3.4: Estimating the global motion, a) two consecutive frames #166 and #167 from road surveillance video 1.1, 50% of each dimension (original 400x640) b) illustration of the displacements using color-coded edge images, c) histogram of the horizontal and vertical displacements, d) motion vectors with an option of using all blocks in the image, normally we can indeed favor the blocks giving a high variance score and eliminate the blocks having ordinary texture.

3.5 Summary

In this chapter, we discussed the task of suppression of the global motion trajectories imposed on the image sequences. Global instabilities are removed using global alignment of the observed frame onto the reference frame, which are simply selected as consecutive frames. Visually, the motion estimation implementation here smooths out the instabilities efficiently. However, the motions are computed using consecutive frames, and a more complex stabilization in which the first and the last frames are forced to come close to each other globally, is not pursued. Instant global displacements due to camera motion are successfully suppressed, leaving out the local displacements due to moving objects. The global instabilities are assumed to be drawing closed paths and not diverging from a zero mean, yet for mobile camera systems neither a simple stabilization nor the estimation of a static background is achievable. After running motion segmentation, local displacements of segmented out regions are localized using a local motion estimation scheme, a well-known method [Thevenaz98] with irregular shaped Region of Support. Not to break up the flow of the document results related to local motion estimation are displayed in Chapter 5.

4 Information Complexity Guided Statistical Background Modeling

Given an image sequence from a fixed image acquisition system separating all foreground objects is the main core of the background subtraction methods. The information coming from the background is suppressed to extract objects in motion. As one of the related research fields, Video Segmentation mostly involves change detection methods to segment each frame into regions, namely the changed and unchanged regions in the case of a stationary imaging system, globally or locally changing regions in the case of non-stationary imaging system. It is mainly used in real-time video applications, such as video surveillance traffic monitoring, and gesture recognition for human-machine interfaces, semantic annotation of videos to name a few [Stauffer99]. It is an integral part of any video analysis and coding problem. Here in this study, video surveillance/content understanding through background suppression is our main concern.

The video data provides the motion differences over time as a strong cue for the moving object segmentation. Motion is the most helpful, yet not the only cue. Observing the same scene with time varying blurs, as well as the variations arising from noise can be utilized to super-resolve a scene. Types of motion analysis used in video related computer vision applications can be generalized basically as follows;

- Motion detection: Detecting any changes in the scene,
- Motion estimation: Localization of the movement of the objects and regions,
- Motion tracking: Correspondence between regions having the same motion behavior,
- Motion recognition: Scenario recognition corresponding to detected motion,
- 3D Structure from motion: Depth related 3D structure of the scene using small camera motions or non-planar object motions.

Motion tracking, motion recognition, structure from motion is totally out of our focus, the former two can represent the practices what we are mainly involved in this study.

Moving object segmentation's main goal when the camera is stationary, is the extraction of objects by change detection and background modeling, and when the camera is moving, compensation of the scene motion first and then executing the same steps. The objective of this chapter is to extract the moving objects at each frame of the video. In contrast, in some applications such as ghost removal in high dynamic range (HDR) images, suppression of motion in between frames and capturing the background is the main objective [Khan06]. Image/Video Reconstruction, in which we combine the information from a set of different versions of the same region of the underlying scene and use it to construct a better representation of the scene with more resolving power, is the further steps after moving object segmentation.

Motion segmentation using background subtraction subdivides an image into its constituent regions or objects and level of division depends on the problem being solved. Precise motion detection/segmentation helps localization of the regions in motion more reliably. Scene complexity is the main factor to choose a specific moving object segmentation method. The amount of camera motion, color and texture uniformity within objects, smoothness of the motion of the objects, objects entering and leaving the scene, slow movement of objects, regularity of the object shape along the temporal dimension, objects overlapping in the visual field, illumination changes due to lighting conditions, moving elements of the scene (e.g. trees,

clouds), and shadows all determine the complexity of the scene. Therefore the real-time video processing applications always limited by computation time and storage barriers. As is often the case, the simplest method is arguably the most robust, on the other hand complex methods take all the factors into consideration at the expense of computational time. More complex scenarios require more sophisticated segmentation algorithms, most of which are either too slow to be practical, or succeeded by restricting themselves to very controlled situations. Also they are impractical for commercial applications. Traditional approaches based on solely background related methods typically fail in complex scenarios given above. Our goal is to create a robust, adaptive scene segmentation system that is flexible enough to handle variations in lighting, moving objects, multiple moving objects and other general arbitrary changes to the observed scene. Situation such as shadows, inter-reflections, objects having the similar characteristics of the background (e.g. the same texture as in camouflage), motionless objects is not in the scope of our study.

A video shot boundary or scene cut detection is the task of finding the instants in a video data that one scene is replaced by another one which is having a different visual content. It is a temporal analysis rather than a spatio-temporal one. A shot is a sequence of frames shot without any interruption and by a single camera. Cut or frame transition detection is mostly required in video indexing, to record the beginning and endings of the shots. Many automatic techniques have been developed to detect transitions in video sequences [Mas03]. In this study rather than indexing the video data into shots, we are involved in spatio-temporal analysis of the video assuming the shots are already in hand.

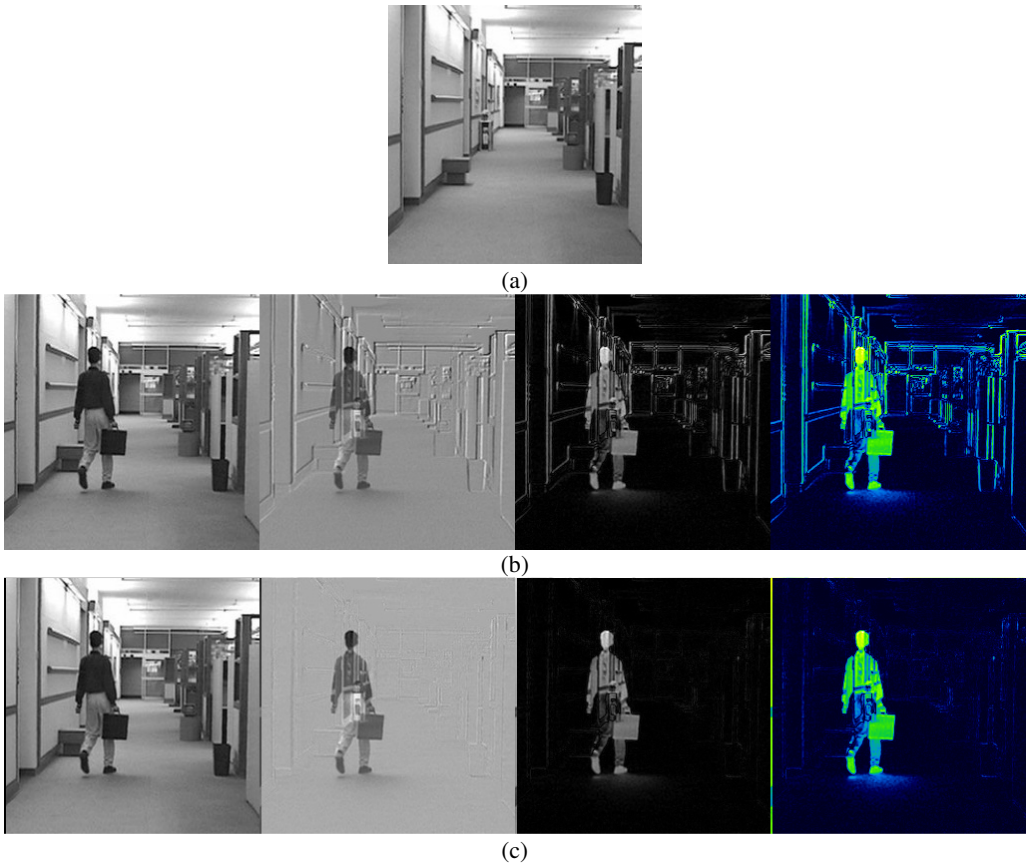
We assume that we are using video data, which is not transmitted or stored in the compressed form, to achieve maximum application independence. Another assumption we make is that image sequence predominantly captures the background, so that in any local region in image space, the number of pixels that capture the background is significantly greater than the pixels that capture the object. Given this assumption, the neighborhood of a pixel in spatio-temporal domain may serve as a reasonable representation of the background.

We are going to discuss moving object segmentation case where the monitoring system or the platform on which the system is mounted on is stationary or the global camera motion is already estimated. If that is not the case instability due to camera platform should be handled, an example of which is given in Figure 4.1.

4.1 Moving Object Detection using Background Subtraction

As the name suggests, in background subtraction two regions should be distinguished; the background which consist of stationary regions of the scene, and the foreground representing the changing regions of the scene over time. Physically, there is no existing method that can accurately find the probability that a pixel is a part of a moving object. As stated in Chapter 2 Background Modeling methods are classified in the following categories by Bouwmans et al. in [Bouwmans10] as; Basic Background Modeling, Statistical Background Modeling, Fuzzy Background Modeling, Neural Network based Background Modeling, Background Modeling by Robust Principal Component Analysis, Background Estimation. In the absence of any a priori knowledge about the target and environment is Basic Background Modeling methods are the most widely adopted, practical approaches for moving object detection in the case of stationary camera system.

The image without moving objects (approximate or precise background) (i) can be chosen as one of the frames, (ii) can be a fixed frame formed using all the frames at once, (iii) or can be initialized and updated for global illumination changes. Subtraction is actually a tool to measure the similarity or dissimilarity between the given frames. The current frame is simply subtracted from the static background, and if the absolute difference in pixel values for a given pixel is greater than a threshold *Th*, the pixel is considered as a part of the foreground:



(a)

(b)

(c)

Figure 4.1: Effect of video stabilization on background subtraction, hallway monitoring a) precise background is known; b) an observed frame from a non-stationary imaging system. Frame difference, absolute frame difference, color coded display are shown, c) instability is removed using alignment of the observed frame onto the reference frame. Frame difference, absolute frame difference, color coded display are shown. Video source: [ww.trace.eas.asu.edu/yuv].

$$FD_k(\mathrm{x}) = |g_k(\mathrm{x}) - b(\mathrm{x})|$$
$$z_k(\mathrm{x}) = \begin{cases} 1 \; if \; FD_k(\mathrm{x}) > Th \\ 0 \; otherwise \end{cases} \tag{4.1}$$

where x denotes pixel location, k is the frame index, $k = 1,2,\ldots,K$, g_k is the observed frame at k^{th} time instant with size mxn for each color channel, b is the background function, z_k is the segmentation label, which equals to '1' for changed regions and '0' for unchanged regions. Since the illumination is more or less constant from one frame to the other one, the values z_k change mostly due to local changes. Here the big challenge is to determine the value *Th,* this is done empirically, however there are ways of adaptively employing it. In some simple cases such as when monitoring a hallway, fewer objects will be detected and the scene stays clear most of the time. After pixel-wise thresholding a process to eliminate the isolated labels is followed. The approximated background as the average or the median of the previous n frames can be adopted if there is no obvious static background in hand. In median filtering K frames of the video are buffered and the background is calculated as the median of the buffered frames. Averaging works the same,

the average of the all pixel values over time are assigned as the value of the corresponding pixel in the background model. A more efficient compromise over median filtering was proposed in [McFarlane95]. Their approximate median method is a more efficient recursive approximation of the median filter. In this method if a pixel in the observed frame has a value larger than the corresponding background pixel, the background pixel is incremented Likewise, if the current pixel is less than the background pixel, the background pixel is decremented. In this way, the background eventually converges to an approximation where half the input pixels are greater than the background, and half are less than the background, approximately the median.

Since the average, median or the approximated median frames are calculated sweeping all the frames over time instants these methods are rather fast but very memory consuming, the memory requirement is K times the size of the frames. Storing and processing many frames of the video requires large amount of memory. A small modification is using consecutive frames:

$$FD_{k,r}(\mathrm{x}) = |g_k(\mathrm{x}) - g_r(\mathrm{x})|$$
$$z_{k,r}(\mathrm{x}) = \begin{cases} 1 \; if \; FD_{k,r}(\mathrm{x}) > Th \\ 0 \quad otherwise \end{cases} \tag{4.2}$$

where g_r denotes the reference frame and for the successive frame difference it can be taken as g_{k-1}. In this case the approximated background is just the previous frame, and if the difference in pixel values for a given pixel is greater than a threshold the pixel is considered as a part of the foreground region. Frame difference based Basic Background Modeling evidently works best if the frame rate is fast enough not to leave big spatio-temporal change gaps between frames. This kind of frame difference analysis is not satisfactory for two reasons first a uniform region may be interpreted as stationary even if it is moving (aperture problem). Second the intensity difference due to motion is affected by the magnitude of the spatial gradient in the direction of motion. The most important advantage of this method is noise suppression, since the background model is based solely on the previous frame, it can adapt to changes in the background faster than any other method (at 1/fps). Another modification is employing running average over basic average frame calculation:

$$b_k(\mathrm{x}) = \alpha . g_{k-1}(\mathrm{x}) + (1-\alpha) . b_{k-1}, \quad b_0 = [0]$$
$$FD_k(\mathrm{x}) = |g_k(\mathrm{x}) - b_k(\mathrm{x})|$$
$$z_k(\mathrm{x}) = \begin{cases} 1 \; if \; FD_k(\mathrm{x}) > Th \\ 0 \quad otherwise \end{cases} \tag{4.3}$$

where b_k denotes the background estimate at k^{th} time instant, $k = 1,2,\ldots,K$, α is the learning rate typically 0.05. As it can be seen easily only the two consecutive frames from the image sequence are used - $g_{k-1}(\mathrm{x})$ and $g_k(\mathrm{x})$. There is no need to store all of the frames. If α equals to 1, then $b_k(\mathrm{x}) = g_{k-1}(\mathrm{x})$ thus it becomes the successive frame difference scheme. An experimental case study for the basic methods described so far is given in Figure 4.3 for the image sequence given in Figure 4.2. In Figure 4.3 (a) the first frame can be used as the background and technically there is no need to use other methods. For the other methods the advantage of having a clear background representation is not used.

For the automatic monitoring systems there is no single frame based initial way to evaluate static regions, the sequence should processed and only the moving regions can be eliminated, and an estimated background model is sought after. The successive frame difference is presented in Figure 4.3(b) the background is taken as the previous frame, the noise suppression is the best however, the frame rate is not high enough thus moving regions are super-imposed. Presented in Figure 4.3(c), Figure 4.4(a-b) the median, average and approximate median background subtraction work also fine because the object occupies a certain region for a very short time, and for the rest of the time that region is unoccupied.

Figure 4.2: An image sequence from a surveillance camera for an experimental case study.

For the complex scenes where there are many moving objects these methods all fail. Figure 4.4 (c) the running average background subtraction is used only to reduce the over memory use and depends heavily on the learning rate α. All of these methods above have low or medium complexity. The detection accuracy can be measured in terms of correctly and incorrectly labeled pixels during normal conditions of the objects motion (stationary background, or fixed camera system). Basic Background Modeling possess high reactivity to immediate start and stop of the objects, and they fit to the only practical uses such as detection of the actual moving objects and elimination transient background changes [Cucchiara03].

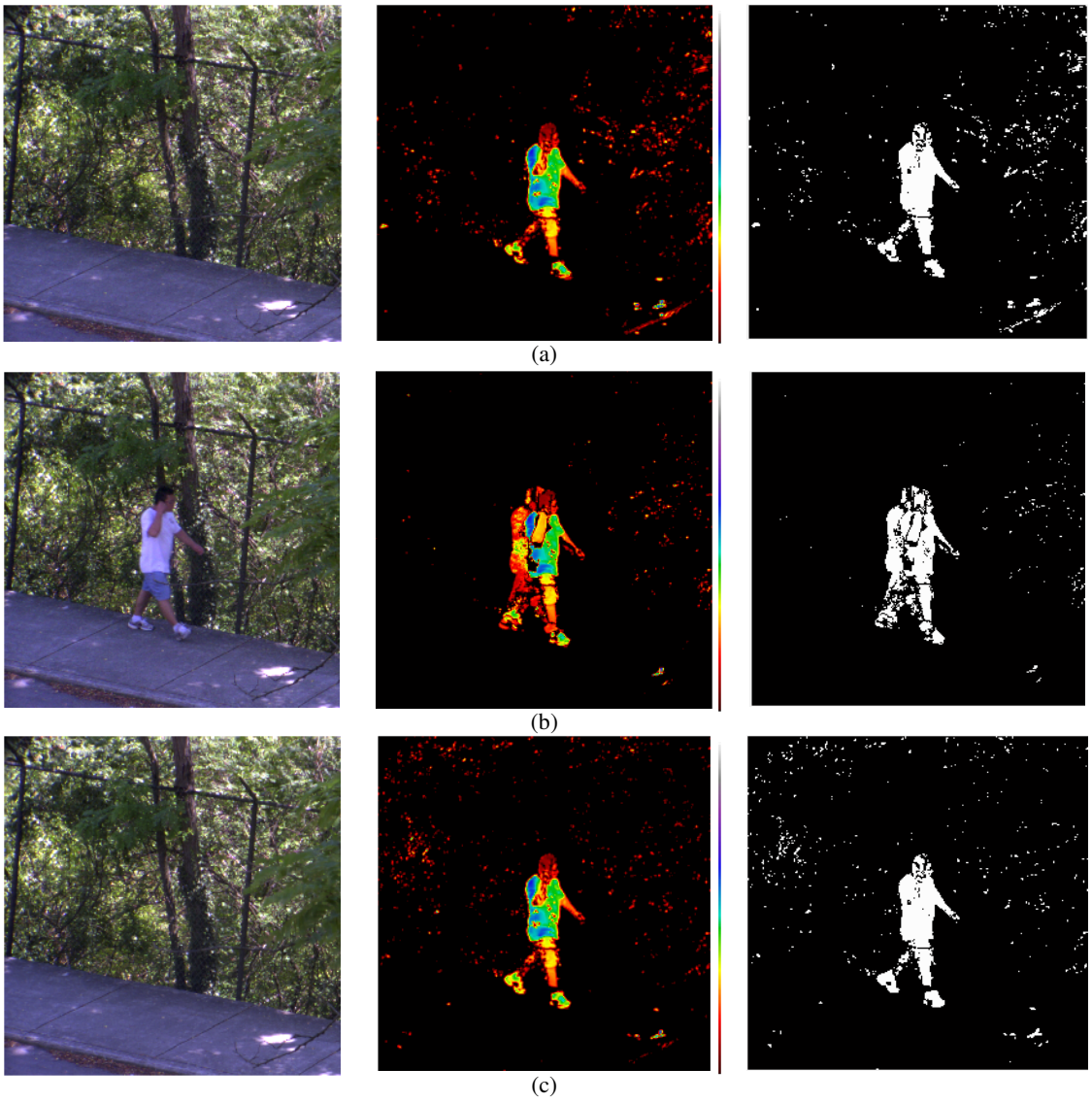


Figure 4.3: Basic foreground estimation using background subtraction for frame 17. Background model, approximated moving regions and segmentation of these regions using a threshold value of 0.07 (intensity range is [0-1]) , for several methods: a) Simple background subtraction, b) Successive background subtraction, c) Median background subtraction. Image size (275x275)

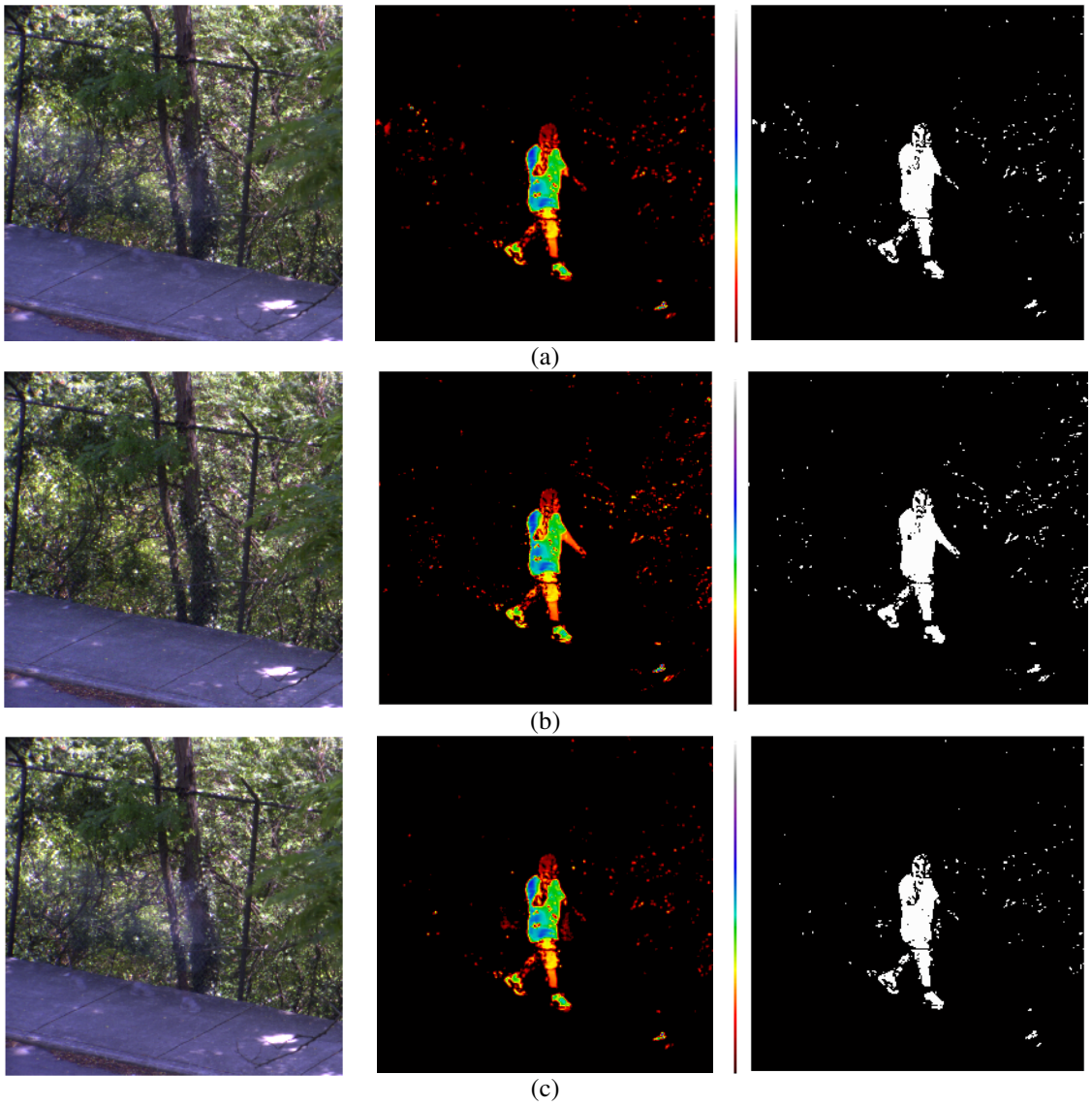


Figure 4.4: Basic foreground estimation using background subtraction for frame 17. Background model, approximated moving regions and segmentation of these regions using a threshold value of 0.07 (intensity range is [0-1]), for several methods: a) Average background subtraction, b) Approximate median background subtraction c) Running average background subtraction for α=0.05.

For Basic Background Modeling no spatial correlation is used between different neighboring pixel locations, operations are aimed to model or to update the background based on each pixel's recent history, and thus considering these assumptions they are 1-D methods along the temporal dimension. They do not provide an explicit method, to choose the threshold method. They cannot handle multi-modal processes such that if a region of the scene is occupied equally with two objects one of them should be regarded as background the other is labeled as moving object.

4.2 Statistical Background Modeling

Physically it is difficult to model the background, and assign weights to pixels accordingly. An alternative to compare pixel values directly is analyzing the statistics of pixels. Is it possible to assign a weight for each pixel that helps us to determine the contribution of each pixel at any time in the image sequence?

Background modeling algorithms have been developed to overcome this problem by modeling and updating the background statistics pixel-wise. Kim et al. classifies such methods into two categories: parametric and non-parametric [Kim07]. The parametric approaches set a background distribution in advance and update the parameters of the model, whereas the non-parametric approaches estimate density function directly using the data. Kim et al.'s non-parametric background modeling is identical to the Background Estimation category of Bouwmans et al. [Bouwmans10]. The most popular distribution is the Gaussian distribution. A single Gaussian to model each pixel's nature is utilized firstly by Wren et al. [Wren97]. The pixels are classified as the elements of either the background or the moving regions. Friedman and Russel in [Friedman97] Stauffer and Grimson in [Stauffer99] published their works on Gaussian Mixture Models (GMMs) based Background Modeling and has become the pioneers of the related studies.

GMMs based Statistical Background Modeling

Pearson [Pearson94] is the first author to model a dataset consists of two populations with a GMM with two Gaussian distribution. For background modeling purposes earlier methods used single Gaussian distribution to model the probability distribution of the pixel intensity [Wren97]. The work of Friedman and Russel [Friedman97] can be regarded the first example of what we conveniently call background modeling and the most similar preceding example to the Stauffer and Grimson's work. They put their efforts for background subtraction for traffic surveillance project. They proposed Mixture of Gaussians (MoG) model based classification for each pixel using an incremental version of Expectation Maximization (EM) to overcome the instabilities of standard time-averaging approaches. The mixture of three Gaussian components are corresponding to road, vehicle and shadows are initialized using an EM algorithm.

Stauffer and Grimson model each pixel as a mixture of Gaussians and use an on-line approximation to update the model. The main assumption is that the video sequences involve i) light changes, ii) scene changes, iii) and moving objects. The Gaussian distributions of the adaptive mixture model are then evaluated to determine which are most likely to result from a background process. Each pixel is classified based on whether the Gaussian distribution which represents it most effectively is considered part of the background model. Authors reported a stable, real-time outdoor tracker reliably dealing with lighting changes, repetitive motions from clutter, and long-term scene changes. These points make the GMMs use a possible deal breaker for many applications. GMMs were proved to be very robust, that they can handle multi-modal distributions. For instance, a leaf waving against a blue sky has two modes—leaf and sky. GMMs can filter out both. Kalman filters [Ridder95] effectively track a single Gaussian, and are therefore unimodal: they can filter out only leaf or sky, but usually not both. In GMMs, normally the background does not consist of single values. Rather, the background model is parametric. The pixel process is mainly a time series of pixel values, scalars for gray values and vectors for color images. Up to a frame number N, the history of the pixel measurements can be represented as $\{X_n | \, n = 1..N\}$. The pixel process can model as follows:

$$P(X_n) = \sum_{k=1}^{K} w_{n,k}\eta(X_n, \mu_{n,k}, \Sigma_{n,k})$$
$$\eta(X_n, \mu_{n,k}, \Sigma_{n,k}) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_{n,k}\right|^{1/2}} e^{-\frac{1}{2}(X_n-\mu_{n,k})^T \Sigma_{n,k}{}^{-1} (X_n-\mu_{n,k})} \quad (4.4)$$

where K is the number of the distributions, $w_{n,k}$ is an estimate of the weight (what portion of the data is accounted for by this Gaussian) of the k^{th} Gaussian in the mixture up to frame number n, $\mu_{n,k}$ is the mean value of the k^{th} Gaussian in the mixture up to frame number n, $\Sigma_{n,k}$ is the covariance matrix of the k^{th} Gaussian in the mixture up to frame number n and η is the Gaussian probability density function. The simplification $\Sigma_{n,k} = \sigma_{n,k}^2\mathbf{I}$ can be adopted assuming that the red, green and blue channels are independent and have the same variances. Their implementation can be divided into 3 parts.

Initialization — The pixel distributions are initialized with the K-means algorithm. The data is clustered into components and the variance, cluster centers and the weights are initialized. EM initialization performs a little better particularly if the weather conditions are dynamic (e.g., fast moving clouds). But, if the area under surveillance were a busy plaza (many moving humans and vehicles), the on-line K-means initialization might have been more preferable [Pavlidis01].

Model Update — When a new value is observed it will be represented by one of the major components and used to update the GMM. Every new pixel value is checked against the existing K Gaussians until a match is found. A match is determined if the pixel value is within a factor of standard deviation of the distribution. This is a kind of adaptive threshold when composing background model. If none of the K Gaussians match the current pixel value the least probable distribution is replaced with a distribution with the current value as its mean value, an initially high variance and low prior weight. The prior weights are updated with the new values:

$$w_{n,k} = (1-\alpha)w_{n-1,k} + \alpha(M_{n,k}) \quad (4.5)$$

$M_{n,k}$ is one for the model which is matched and zero for the remaining models, α is the learning rate. After this update al weights are normalized. The μ and σ parameters remain the same for unmatched distributions. For the matched distribution, parameters are updated as follows:

$$\mu_{n,k} = (1-\rho)\mu_{n-1,k} + \rho X_n\,, \quad \rho = \alpha\,\eta(X_n, \mu_{n,k}, \Sigma_{n,k})$$
$$\sigma_{n,k}^2 = (1-\rho)\sigma_{n-1,k}^2 + \rho(X_n, -\mu_{n,k})^T(X_n - \mu_{n,k}) \quad (4.6)$$

Background composition — The weight and standard deviations of each component are measures of the confidence in the pixel value guess (higher weight & lower standard variation, $w/\sigma \rightarrow$ higher confidence). After the process to determine if a pixel is part of the background, a comparison takes place. If the pixel value is within a scaling factor of a background component's standard deviation σ, it is considered part of the background. Otherwise, it's foreground. First, the Gaussians are ordered by the value of w/σ. This value increases both as the distribution gains more evidence and as the variance decreases. This ordering of the model is effectively an ordered, open-ended list, where the most likely background distributions remain on top and the less probable transient background distributions gravitate towards the bottom. After establishing the GMMs for all the pixels an image is composed the means of most probable components related to each pixel, and foreground pixels are segmented out.

For the sake of running an online fast responding background modeling system, important temporal and spatial constraints are not tackled. Yet following discussion is also given in Chapter 2, we want to mention about the same issues to clarify to idea of our contribution in this study. The most important assumptions which are extended by many others were;

– The number of clusters K are blindly assigned,

- Covariance matrix for 3-D (pixel values from RGB space) Gaussian components at a time instant has the form $\sigma_k^2\mathbf{I}$, where σ_k^2 are variances for each channel, and $\mathbf{I}$ is the identity matrix. Namely for any component the channels are assumed to be independent from each other. But these color components are surely dependent and so the simplification made here for the covariance matrix is not right.
- Online K-means approximations instead of EM based more robust component update scheme is used,
- This algorithm does not distinguish shadows from objects and the normalized color module fails when the input signal has no color and its discrimination power is poor in dark and saturated areas.
- Another disadvantage, which is a general for background modeling using a pixel-wise aspect, is that it prevents to handle some critical situations which can be only detected spatially and temporally.
- Furthermore, some critical situations need pre-processing or post-processing (e.g. camera motion).

Such critical situations, as well as the real time constraints are investigated by many others and there is a good list of these methods in [Bouwmans10]. We want to continue the discussion on the number of Gaussian components K in the context of Information Complexity guided Statistical Background Modeling. An optimal selection of number of components K as well as the shape of the components will also be given in the sub-chapters.

4.3 Optimal Number and Shape of Gaussian Components in GMMs Based Background Modeling

Clustering or classification is a widely used task in various fields of science. We can regard clustering in general as the automated tools to establish categorization based on a criterion (such as similarity) imposed on the measurements, findings or concepts. When there is no prior information on the grouping of the data, unsupervised learning tools, in other words clustering methods are required. An approximated order on the complexity of data is obtained through clustering.

Clustering methods try to imitate what the human vision-evaluation system does well in low dimensions. Beyond two dimensions however, and even in some two dimensional cases, understanding the complexity of things and discovering an order within that complexity, becomes a problem that is still lacking a widely accepted solution. There are various techniques studied in cluster analysis literature which methods can be divided into two basic types: hierarchical and partitional clustering [Jain88], yet there exists a number of subtypes and different algorithms for clustering. *Hierarchical clustering* — either merges smaller clusters into larger ones, or splits larger clusters. Merging process is named as agglomerative and splitting is named as divisive. At the end of the algorithm is a tree of clusters -a dendrogram- is obtained, which shows how the clusters are related. At a m level a clustering of the data dendrogram is cut and separated groups are obtained. *Partitional clustering* — on the other hand, attempts to directly decompose the dataset into a set of separated clusters. The criterion function tries to minimize the global structure of the data distribution through utilizing the measure of dissimilarity in the samples within each cluster, while maximizing the dissimilarity of different clusters. A commonly used partitional clustering method is K-means clustering which partitions the data into K groups by minimizing the within-group sum of squares. In K-means clustering, the criterion function is the averaged square distance of the data points from their nearest centroids where the sum of distances between each point and the closest class center to the point should be minimized.

The factors such as; the shape and separation of clusters, similarity of shape from one cluster to another, relative sizes and compactness of clusters, dimensionality and the number of observations makes the clustering problem specific to the scenario in hand. A problem with the clustering methods is that the interpretation of the clusters may be difficult. Most clustering algorithms prefer certain cluster shapes, and the algorithms will always assign the data to clusters of such shapes even if there were no clusters in the

data. Therefore, if the goal is not just to reduce the dimensionality of the dataset but also to make inferences about its cluster structure, it is essential to analyze whether the dataset reveals a clustering tendency. Furthermore, it can be difficult to tell from data whether a physical or other observed process is random or chaotic. Trying to cluster a chaotic data is not possible. There will always be some form of corrupting noise, even if it is present as round-off or truncation error. Thus any real physical data series, even if mostly deterministic, will contain some randomness. A deterministic system will have an error that either remains small (stable, regular solution) or increases exponentially with time (chaos). A stochastic system will have a randomly distributed error. Behavior of image acquisition process in a time period can be regarded as stochastic rather than chaotic. We do not expect exponential changes for any pixel.

Assuming the data has an acceptable clustering tendency, another potential problem is the choice of the number of clusters. It may be critical; quite different kinds of clusters may emerge when K is changed. Good initialization of the cluster parameters (a number of parameters depends on the model adopted) is crucial. Some clusters will be empty during update scheme if initially assigned centroids lie far from the dense regions of the data. Clustering can be used both to reduce the dimensionality of data and to satisfy a good categorization. After clustering it may not be still obvious what the outcome is. The clusters should be shown somehow to give an idea to the user about what they are like, thus some additional means are needed for visualizing them.

4.3.1 Model Based Clustering

In model based clustering, a collection of concepts and quantities forms a solid ground for examining the grouping structure in a dataset. The group memberships are learned, maintained and updated parametrically. The presence of multi-variate data in many applications motivates researchers to use statistical tools rather more frequently since the computer technology help us with its heavy processing capability. The clusters consist of N points in a p-dimensional space are assumed to be coming from K different populations. Generating a standard statistical model requires handling with a mixture of K underlying populations, each of which is a cluster. Therefore we can regard our clustering problem now transformed into a parameter estimation problem. The determination of such points;

- the form of components,
- the number of components,
- an optimization method for clustering using a certain form and number of components,
- criteria to determine the optimal model, evaluation of the information complexities for a number of model options,

is required for mixture model based clustering [Erar2011]. There are numerous distributions to use as density components, as well as a vast number of different optimization methods and model selection criteria to decide on best fitting model. Since a number of options can be considered for the solution, there is a significant amount of opportunity available in the development of the method.

4.3.2 Gaussian Mixture Models

In 1809, Carl Friedrich Gauss introduced the theory of the Normal distribution which is also called the Gaussian distribution, after him. A great number of researchers have adopted Gaussians as the distribution to model the data clusters.

GMM can be used for cluster analysis. The researchers when each pixel history $X \in \mathbb{R}^{(Nxp)}$ is given (p-dimensional N points), would be interested in estimating the number of populations (a.k.a groups, clusters, or classes) K. The class membership of each observation in the history will be ($\hat{y}_n \mid X, n = 1 \dots N,\ \hat{y}_n \in \{1, \dots, K\}$) based on posteriori of the data. The hat of $\hat{y}_n$ indicates it is an estimation not a precise classification. The GMM, in this case, is a useful tool to the researchers which helps to fit a mixture probability density function to the given data; also it allows implementation of other formal statistical

procedures for estimation and optimization. Assuming the p-dimensional observations $x_n|n = 1 \dots N$ come from a mixture of K underlying probability distributions, each corresponding to a different cluster, the mixture density will be given by:

$$f(\mathrm{x};\mathrm{w},\theta) = \sum_{k=1}^{K} w_k \eta\,(\mathrm{x};\theta_k) \tag{4.7}$$

where w_k is mixing proportions and satisfies $w_k > 0$, $\sum_{k=1}^{K} w_k = 1$, θ_k is the vector of unknown parameters of the k^{th} component, and η represents the probability that an observation belongs to the k^{th} component. Multi-variate η function with parameters $\theta_k = (\mu_k, \Sigma_k)$ is:

$$\eta(\mathrm{x};\ \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{p/2}\,|\Sigma_k|^{1/2}}\ \exp\{-\frac{1}{2}(\mathrm{x}-\mu_k)^T\Sigma_k^{-1}(\mathrm{x}-\mu_k)\} \tag{4.8}$$

Approaching the clustering problem from this probabilistic standpoint reduces the whole problem to the parameter estimation of a mixture density. The unknown parameters of the Gaussian mixture density given in Equation 4.8 are the mixing proportions w_k, the mean vectors μ_k, and the covariance matrices Σ_k. Therefore, to estimate these parameters, we need to maximize our confidence to the parameter estimation. Since the data is the realization of repeated experiments drawn independently likelihood function of parameters for a single component weighted with mixing proportion is:

$$\begin{aligned} &f(x_n; w_k, \theta_k) = w_k \eta(x_n; \theta_k), \\ &L(\theta_k|\mathrm{X}) = f(x_1, x_2, \dots, x_N; w_k, \theta_k) = f(x_1; w_k, \theta_k) f(x_2; w_k, \theta_k) \dots f(x_N; w_k, \theta_k) \\ &= \prod_{n=1}^{N} w_k\, \eta(x_n; \mu_k, \Sigma_k) \end{aligned} \tag{4.9}$$

Taking the logarithm of the term given in Equation 4.9 gives the log-likelihood related to a single component and since we have K components summing the up is the Gaussian mixture model's log-likelihood:

$$\begin{aligned} \log L(\theta_k|\mathrm{X}) &= \sum_{n=1}^{N} \log\{w_k \eta(x_n; \theta_k)\} \\ \log L(\theta|\mathrm{X}) &= \sum_{k=1}^{K}\sum_{n=1}^{N} \log\{w_k \eta(x_n; \mu_k, \Sigma_k)\} \end{aligned} \tag{4.10}$$

There is no closed form solution to $\log L(\theta|\mathrm{X}) = 0$ for any distribution mixture; so the likelihood has to be maximized numerically. For this numerical optimization, the Expectation-Maximization (EM) algorithm is used widely.

EM as a Numerical Optimization for GMM Based Clustering

The EM algorithm is an iterative procedure consisting of two alternating steps, given some starting values for the parameters given as:

$$\widehat{w}_k = \frac{1}{N}\sum_{n=1}^{N} Cl_k(\hat{y}_n), \quad \hat{\mu}_k = \frac{1}{\widehat{w}_k N}\sum_{n=1}^{N} x_n Cl_k(\hat{y}_n)$$

$$\hat{\Sigma}_k = \frac{1}{\widehat{w}_k N}\sum_{n=1}^{N}[(x_n - \hat{\mu}_k)^T(x_n - \hat{\mu}_k)]Cl_k(\hat{y}_n) \tag{4.11}$$

$$Cl_k(\hat{y}_n) = \begin{cases} 1 & \hat{y}_n = k \\ 0 & \hat{y}_n \neq k \end{cases}$$

The EM algorithm is an iterative procedure consisting of two alternating steps, given some starting values for the parameters in Equation 4.11. $Cl_k(\hat{y}_n)$ values are computed via an initialization scheme (i.e. K-means clustering). Iterative steps of EM algorithm can be summarized as follows:

1. Start with the initial parameters $\widehat{w}_k^{(t)}, \hat{\mu}_k^{(t)}, \hat{\Sigma}_k^{(t)}|_{t=0}$
2. Apply *E-step* and *M-Step* iteratively for $t = 0 \ldots T$ or until a stopping criteria met:

   $(t \geq T)$ && $\left(\left|\left(logL^{(t+1)}(\hat{\theta} \mid \mathrm{X}) - logL^{(t)}(\hat{\theta} \mid \mathrm{X})\right)/\left(logL^{(iter)}(\hat{\theta} \mid \mathrm{X})\right)\right| > Ltol\right)$

   2.1. *E-step* —the posterior probability $\hat{\tau}_{ik}$ of the n^{th} observation belonging to k^{th} component is estimated employing the previous parameter estimates:

$$\hat{\tau}_{n,k} = \frac{\widehat{w}_k^{(t)}\eta\left(x_n; \hat{\mu}_k^{(t)}, \hat{\Sigma}_k^{(t)}\right)}{\sum_{k=1}^{K}\widehat{w}_k^{(t)}\eta\left(x_n; \hat{\mu}_k^{(t)}, \hat{\Sigma}_k^{(t)}\right)} \tag{4.12}$$

   2.2. *M-step* — the parameter estimates of w_k, μ_k, and Σ_k are updated given the estimated posterior probabilities:

$$\widehat{w}_k^{(t+1)} = \frac{1}{N}\sum_{n=1}^{N}\hat{\tau}_{n,k}$$

$$\hat{\mu}_k^{(t+1)} = \frac{1}{N\widehat{w}_k^{(t+1)}}\sum_{n=1}^{N} x_n\hat{\tau}_{n,k} \tag{4.13}$$

$$\hat{\Sigma}_k^{(t+1)} = \frac{1}{N\widehat{w}_k^{(t+1)}}\sum_{n=1}^{N}[(x_n - \hat{\mu}_k^{(t+1)})^T(x_n - \hat{\mu}_k^{(t+1)})]\hat{\tau}_{n,k}$$

There are important issues to be addressed in the EM algorithm:

- Determining the number of components *K*,
- Initialization of the parameters $\widehat{w}_k^{(t)}, \hat{\mu}_k^{(t)}, \hat{\Sigma}_k^{(t)}$,
- Different structures for the covariance matrices, which will lead to different update equations for the covariance matrix simpler than $\hat{\Sigma}_k^{(t+1)}$ computations of Equation 4.13,

Among the GMMs having different covariance structures and number of components, an evaluation should be used to find which model fits the best. Mostly clustering techniques use empirical and subjective means selection of the number of clusters. The most common procedure used in the literature is to fit different models to a range of cluster numbers, $k = 1, 2, \ldots, K$, and then picking up the best fitting model. Neither determining initial parameter values to pass on to the EM algorithm nor experimenting on the number of clusters is easy. In the literature, other less computationally intensive clustering methods are usually used

for the initialization. As stated previously the famous K-means algorithm, is an option. Despite the fact that it has wide use in the literature, the K-means algorithm has its shortcomings as a clustering algorithm itself. One obvious disadvantage is the necessity of initial values to start the K-means algorithm itself and it is not robust to the selection of these initial values. Model-based hierarchical agglomerative clustering is a good alternative to K-means to initialize the EM algorithm. It is a hierarchical clustering algorithm; the clusters are merged by maximizing the classification likelihood. However for our purpose it is computationally very expensive. Despite K-means' shortcomings we prefer using it as the initialization scheme for EM which iteratively updates parameters for an efficient clustering.

Different Structures of the Covariance Matrices

The GMM given in Equation 4.8 assumes the covariance matrix of each component is different, or in other words, there is no simplification on the covariance matrices. One obvious disadvantage of using this general model is that the maximum number of parameters to represent the covariance matrix has to be sought, and each additional parameter indicates an increase in the computational time subject to the size of the dataset. Using this model is against the principle of parsimony. This actually is a more important concern from the viewpoint of expediency.

The covariance matrices in general represent the geometric features, namely, volume, shape and orientation of the clusters. All these geometric features are different for each cluster of the GMM if utilizing a general, complex form of the covariance matrix. However mostly these features are simpler than we assume. Therefore, if simpler or more suitable models for the covariance matrices are derived, such models bring parsimony into the clustering. If the user has an insight about the structure of the covariance matrix, a certain type can be imposed. For such a purpose simpler and easily interpretable parameterized models were established by Banfield and Raftery [Banfield93]. The geometric features of the clusters can be distinguished using eigen-value decomposition of the covariance matrix. The eigenvalue decomposition of the k^{th} covariance matrix is given as $\Sigma_k = \lambda_k D_k A_k D_k^T$ where λ_k is a scalar, D_k is the orthogonal matrix of eigenvectors and A_k is a diagonal matrix containing the normalized eigenvalues, such that $|A_k|$=1. The volume of the cluster is specified by λ_k, which is proportional to the volume of the standard deviation ellipsoid; D_k determines the orientation of the cluster while A_k is associated with the shape of the density. To construct a GMM with $\Sigma_k = \lambda_k \mathrm{D}_k \mathrm{A}_k \mathrm{D}_k^T$, μ_k, w_k, $K-1$ parameters for weights, Kp parameters for means, $Kp(p+1)/2$ (diagonal elements and upper or lower elements of all Σ_k models) for the covariance matrix overall $m = Kp + K - 1 + Kp(p+1)/2$ parameters are needed (K is the number of the components, p is the dimension of the data). For means and weights it is always $\alpha = Kp + K - 1$ parameters. For each $\Sigma_k = \lambda_k \mathrm{D}_k \mathrm{A}_k \mathrm{D}_k^T$ type covariance matrix $Kp(p+1)/2$ parameters overall $m = \alpha + Kp(p+1)/2$ parameters are needed. For $\Sigma_k = \lambda_k \mathrm{I}$ type covariance matrix $m = \alpha + K$ parameters are needed similar taxonomy can be used for other model structures given in Table 4.1. Celeux and Govaert [Celeux95] give the definitions and derivations of all 14 available models, along with the covariance matrix update equation to be evaluated in the M-step of EM algorithm. Nine of these models that have closed form solutions to the covariance matrix update equation can be used easily. A brief summary of descriptions to these models are given in Table 4.1.

Information Complexity Criteria for Model Selection

The covariance models described above in Table 4.1 embraces the geometric features of cluster densities differently; also they require different derivation and update schemes when used in EM like iterative algorithms. Our assumption is that we are not acknowledged about the number of clusters in GMMs based Background Modeling. Additionally, the optimal covariance structure should be determined simultaneously with the optimal number of clusters and one strictly depends on the other. Thus, setting different combinations, which leads to different distribution models, is inevitable.

Table 4.1: Parameterizations of the covariance matrix and the corresponding geometric features*

ID	Volume	Shape	Orientation	**Covariance Decomposition	Number of Parameters
EII	Equal	Equal	N/A	$\lambda \mathrm{I}$	$\alpha + 1$
VII	Variable	Equal	N/A	$\lambda_k \mathrm{I}$	$\alpha + K$
EEI	Equal	Equal	Axes	$\lambda \mathrm{B}$	$\alpha + p$
EVI	Equal	Variable	Axes	$\lambda \mathrm{B}_k$	$\alpha + Kp - K + 1$
VVI	Variable	Variable	Axes	$\lambda_k \mathrm{B}_k$	$\alpha + Kp$
EEE	Equal	Equal	Equal	$\lambda \mathrm{DAD}^{\mathrm{T}}$	$\alpha + p(p+1)/2$
EEV	Equal	Equal	Variable	$\lambda \mathrm{D}_k \mathrm{AD}_k^T$	$\alpha + Kp(p+1)/2 - (K-1)p$
EVV	Equal	Variable	Variable	$\lambda \mathrm{D}_k \mathrm{A}_k \mathrm{D}_k^T$	$\alpha + Kp(p+1)/2 - (K-1)$
VVV	Variable	Variable	Variable	$\lambda_k \mathrm{D}_k \mathrm{A}_k \mathrm{D}_k^T$	$\alpha + Kp(p+1)/2$

*The models here have a closed form solution to covariance matrix update equation to be evaluated in the M-step of the EM algorithm. ** $\{\lambda, \lambda_k\}$ is a scalar, I is the identity matrix, $\{\mathrm{B}, \mathrm{B}_k\}$ is a diagonal matrix, and $\{|\mathrm{B}|, |\mathrm{B}_k|\} = 1, \{\mathrm{D}, \mathrm{D}_k\}$ is the orthogonal matrix of eigenvectors and $\{\mathrm{A}, \mathrm{A}_k\}$ is a diagonal matrix containing the normalized eigenvalues, such that $\{|\mathrm{A}|, |\mathrm{A}_k|\} = 1$.

After estimating the parameters for each given combination, the last step is determination of the optimal cluster structure which is the outcome of the best fitting model as the clustering solution [Erar2011]. Two kinds of schemes can be followed to determine which the best is; heuristic approaches such as cross-validation and theoretical approaches such as likelihood weighted information criteria. We will continue our discussion on information criteria.

Schwarz's Bayesian Criterion (SBC) also known as Bayesian Information Criterion (BIC) [Schwarz78] is the most widely used one in model-based clustering studies. Other well-known criterion, Akaike's Information Criterion (AIC) [Akaike73] is a preceding theory of SBC. In this study, we mainly adopted the studies of Bozdogan [Bozdogan94, Bozdogan10] in order to obtain GMMs which are optimal according to information complexity methodology. Additional to these two variances of his information complexity criterion (ICOMP), SBC and AIC criteria are used for comparison. In the rest of this section we will go over these information criteria.

For a general multi-variate model, the loss function can be defined using the terms; likelihood of the model given the parameters (lack or degree of fit), the complexity of having too many model parameters (lack of parsimony), the complexity of model errors (profusion of complexity) [Bozdogan94]. The model giving the lowest score w.r.t an information criterion provides the best balance between good fit and parsimony. In both AIC and SBC only the first two terms are penalized:

$$AIC = -2\, log\, L\left(\hat{\theta} \mid \mathrm{X}\right) + 2m \tag{4.14}$$

$$SBC = -2\, log\, L\left(\hat{\theta} \mid \mathrm{X}\right) + mlog(N) \tag{4.15}$$

m is the number of independent parameters to be estimated and $\hat{\theta}$ is the maximum likelihood estimate for parameter θ, N is the number of data points. In both equations $-2\, log\, L\left(\hat{\theta} \mid X\right)$ is the bad fitting model penalty, which is negative twice the maximized log likelihood. The difference is in the penalty term for model complexity. The lack of parsimony is penalized in terms of the number of parameters. They both trade off a good fit to the dataset with the desire to use as few parameters as possible. If $log(N) > 2$ or $N \geq 8$ it is obvious that AIC penalizes the number of parameters more than SBC does.
ICOMP criterion was proposed by Bozdogan. Lack of fit of a model is penalized by twice the negative of the maximized log-likelihood which is identical to the same first term of AIC and SBC. However, in

ICOMP, a combination of the term for lack of parsimony and a novel term for profusion of complexity are also simultaneously penalized by a scalar complexity measure C, which is function of the model covariance matrix. Using ICOMP under-fitting and the over-fitting of the model could be well-balanced when different combinations are set. ICOMP is defined as:

$$ICOMP = -2\,log\,L(\hat{\theta} \mid X) + 2C(\hat{\Sigma}_{model}) \tag{4.16}$$

where $L(\hat{\theta} \mid X)$ is the maximized likelihood function, C is a real-valued complexity measure and $\hat{\Sigma}_{model}$ represents the estimated covariance matrix of the parameter vector of the model. The covariance matrix is estimated by the inverse Fisher information matrix (IFIM), $\mathcal{F}^{-1}$. In Equation 4.16 the first component of ICOMP measures the lack of fit of the model and the second component measures the complexity of the estimated IFIM. The first order maximal entropic complexity can be defined as:

$$C(\hat{\Sigma}_{model}) = C_1(\hat{\mathcal{F}}^{-1}) = \frac{s}{2} log\left[tr(\hat{\mathcal{F}}^{-1})/s\right] - \frac{1}{2} log\left|\hat{\mathcal{F}}^{-1}\right| \tag{4.17}$$

where $s = \dim(\hat{\mathcal{F}}^{-1}) = rank(\hat{\mathcal{F}}^{-1})$, $\hat{\mathcal{F}}^{-1}$ is the inverse Fisher information matrix. The general form of ICOMP using IFIM is;

$$ICOMP_{IFIM} = -2\,log\,L(\hat{\theta} \mid X) + 2C_1(\hat{\mathcal{F}}^{-1}) \tag{4.18}$$

Another form of ICOMP can be derived as a Bayesian criterion close to maximizing a posterior expected utility (PEU). It is obtained by combining two utility functions; one relating to the lack of fit term, which estimates the KL information, and the other relating to the complexity of the model in terms of the inverse-Fisher information matrix of the parameter manifold of the fitted models. $ICOMP_{PEU}$ can be computed as [Bozdogan10]:

$$ICOMP_{PEU} = -2\,log\,L(\hat{\theta} \mid X) + m + log(N)C_1(\hat{\mathcal{F}}^{-1}) \tag{4.19}$$

For all the criteria discussed here, the decision rule is to select the model that gives the minimum score for the loss function. Computation of ICOMP for the Gaussian mixture model requires the derivation of the inverse Fisher information matrix (IFIM), which is given by [Bozdogan94]. After some simplification, it appears that calculation of the IFIM itself is not necessary for this computation. Using only the traces and determinants of the component covariance matrices, ICOMP for the Gaussian mixture model can be computed easily as:

$$\begin{aligned} ICOMP(\hat{\mathcal{F}}^{-1}) &= -2logL(\theta|X) + 2C_1(\hat{\mathcal{F}}^{-1}) \\ 2C_1(\hat{\mathcal{F}}^{-1}) &= m(log\left[\sum_{k=1}^{\hat{K}}\left\{\frac{tr(\hat{\Sigma}_k)}{\hat{w}_k} + \frac{1}{2}\left(tr(\hat{\Sigma}_k^2) + tr(\hat{\Sigma}_k)^2 + 2\sum_{j=1}^{p}(\sigma_{k,jj}^2)^2\right)\right\}\right] - logm) \\ &\quad - [(p+2)\sum_{k=1}^{\hat{K}} log|\hat{\Sigma}_k| - p\sum_{k=1}^{\hat{K}} log(\hat{w}_k N)] - \hat{K}plog(2N) \end{aligned} \tag{4.20}$$

Where $\sigma_{k,jj}^2$ represents the j^{th} diagonal element of $\hat{\Sigma}_k^2$ and m is the number of parameters corresponding to a given covariance model from Table 4.1. $ICOMP_{PEU}$ will be just multiplying the remaining terms coming after $-2logL(\theta|X)$ with $log(N)/2$ and adding m.

4.4 Experimental Results

We apply the GMM based clustering to a sample dataset and our real world video data for background modeling purpose. As we apply the method to the datasets, we compute scores of various model selection

criteria discussed previously. We evaluate the performance of these different criteria based on clustering using combinations of $i = 1..i_{max} | i_{max} \leq 9$ covariance models and $k = 1 \ldots k_{max}$ clusters. Overall, steps of GMMs based clustering of a general dataset is given it the flow diagram in Figure 4.5.

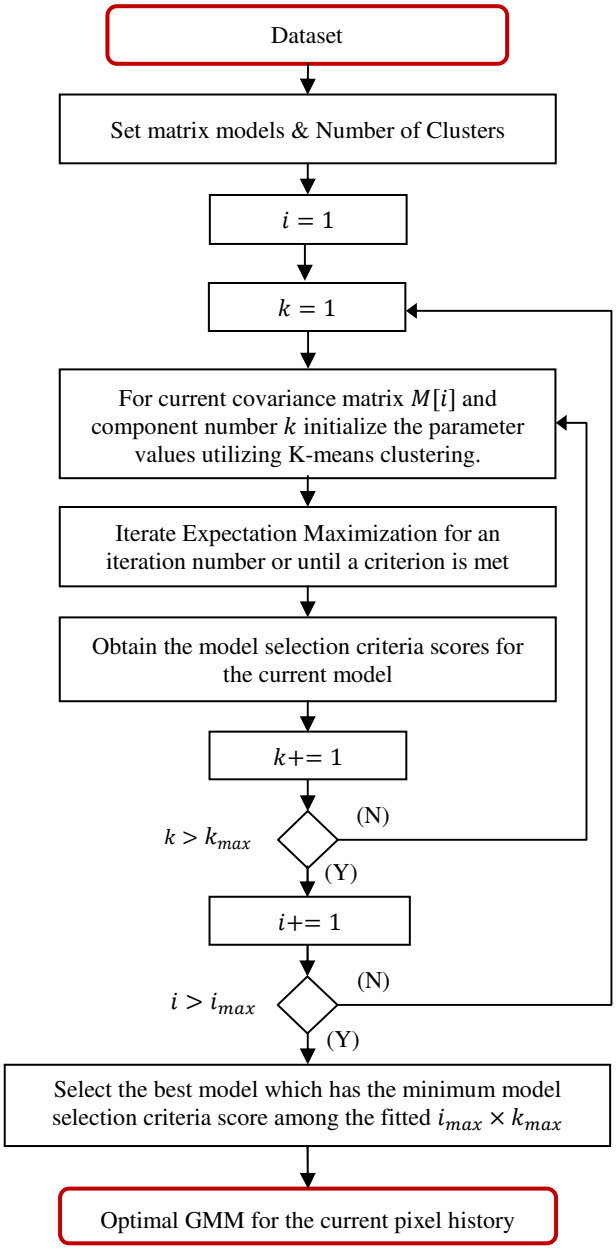


Figure 4.5: Steps of GMMs based clustering.

The maximum for the number of clusters, k_{max} (Figure 4.5) has to be determined. An empirical formula was suggested by Bozdogan [Bozdogan94] to determine the maximum number of clusters, k_{max}:

$$\left(k_L = O\left[\frac{N}{\log N}\right]^{1/3}\right) < \left(k_{max} = N^{0.3}\right) < \left(k_U = \left(\frac{N}{2}\right)^{1/2}\right) \tag{4.21}$$

where $k_L$and k_U are the lower and upper bounds of the maximum number of clusters, $O[\cdot]$ is the order of, and N is the number of observations.

Results of the Experimental Study of Clustering on a Sample Dataset

A bivariate dataset generated from the unconstrained model (Model [VVV]) with 4 groups are used. The group sizes are $\mathrm{N}_1 = 200, \mathrm{N}_2 = 150$ and $\mathrm{N}_3 = 100, \mathrm{N}_4 = 75$. The groups are overlapping and all geometric features vary between groups. Sample scatterplots of the data are shown in Figure 4.6 with labeled clusters. Also in Figure 4.7 surface plot of the mixture of Gaussian components is given. The parameters related to the created Gaussians are given in Table 4.2.

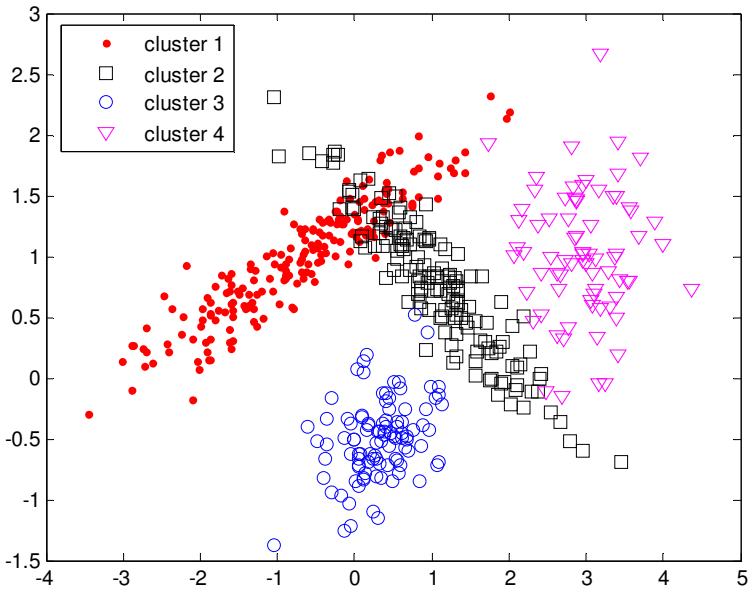


Figure 4.6: Scatterplot of the dataset with labels for the synthetic data.

Table 4.2: Gaussian components' parameters for the synthetic data.

Component	Mean Value	Covariance Matrix	Weight
1	$[-0.7\ 1]$	$\begin{bmatrix}1.2 & 0.5\\0.5 & 0.25\end{bmatrix}$	0.3810
2	[1 0.8]	$\begin{bmatrix}0.5 & -0.35\\-0.35 & 0.3\end{bmatrix}$	0.2857
3	[0.3 –0.5]	$\begin{bmatrix}0.15 & 0.05\\0.05 & 0.1\end{bmatrix}$	0.1905
4	[3 1];	$\begin{bmatrix}0.2089 & 0.0223\\0.0223 & 0.2560\end{bmatrix}$	0.1905

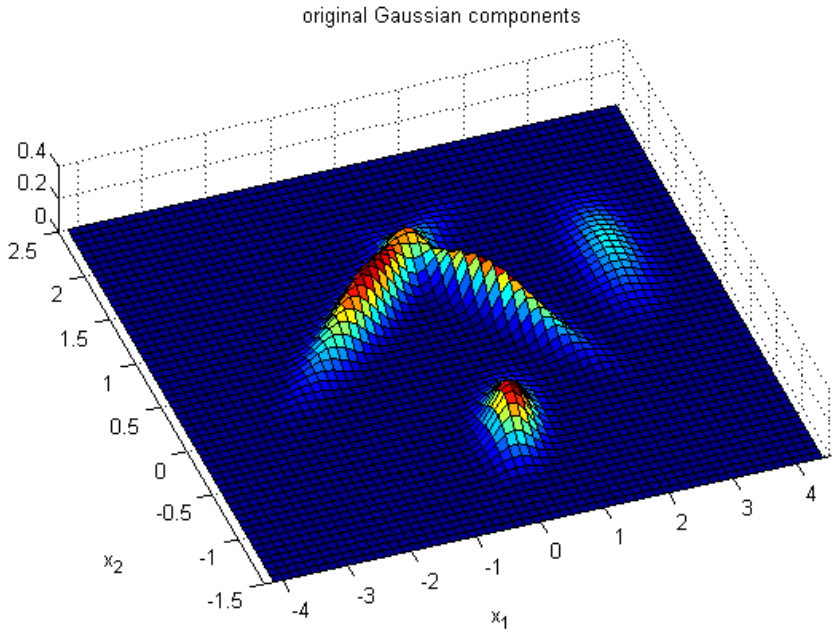


Figure 4.7: Surface plot of the mixture density for the synthetic data.

For a number of clusters k and a covariance model we initialize the parameter values using K-means clustering. The EM algorithm is run for a number of iterations or until a stopping criterion is met. Estimated Gaussians at some iteration steps just for one combination with $K = 4$ and covariance form [VVV] is projected onto the data and displayed in Figure 4.8. We repeated parameter initialization, EM runs for different numbers of n component and covariance forms, and obtain the model selection criteria scores for each model. The model selection results for all four criteria scored, namely AIC, SBC, $\text{ICOMP}_{\text{IFIM}}$ or ICOMP and $\text{ICOMP}_{\text{PEU}}$ are given in Figure 4.9. Recall that the true covariance model here is the unconstrained model [VVV] with $K = 4$ clusters. $\text{ICOMP}_{\text{PEU}}$ gives more importance to the correct K number when compared to other criteria, thus giving a decision based on $\text{ICOMP}_{\text{PEU}}$ results seems to be more reasonable. AIC and $\text{ICOMP}_{\text{IFIM}}$ tend to overestimate the number of components, while SBC tends to select the good model but it is not better than $\text{ICOMP}_{\text{PEU}}$.

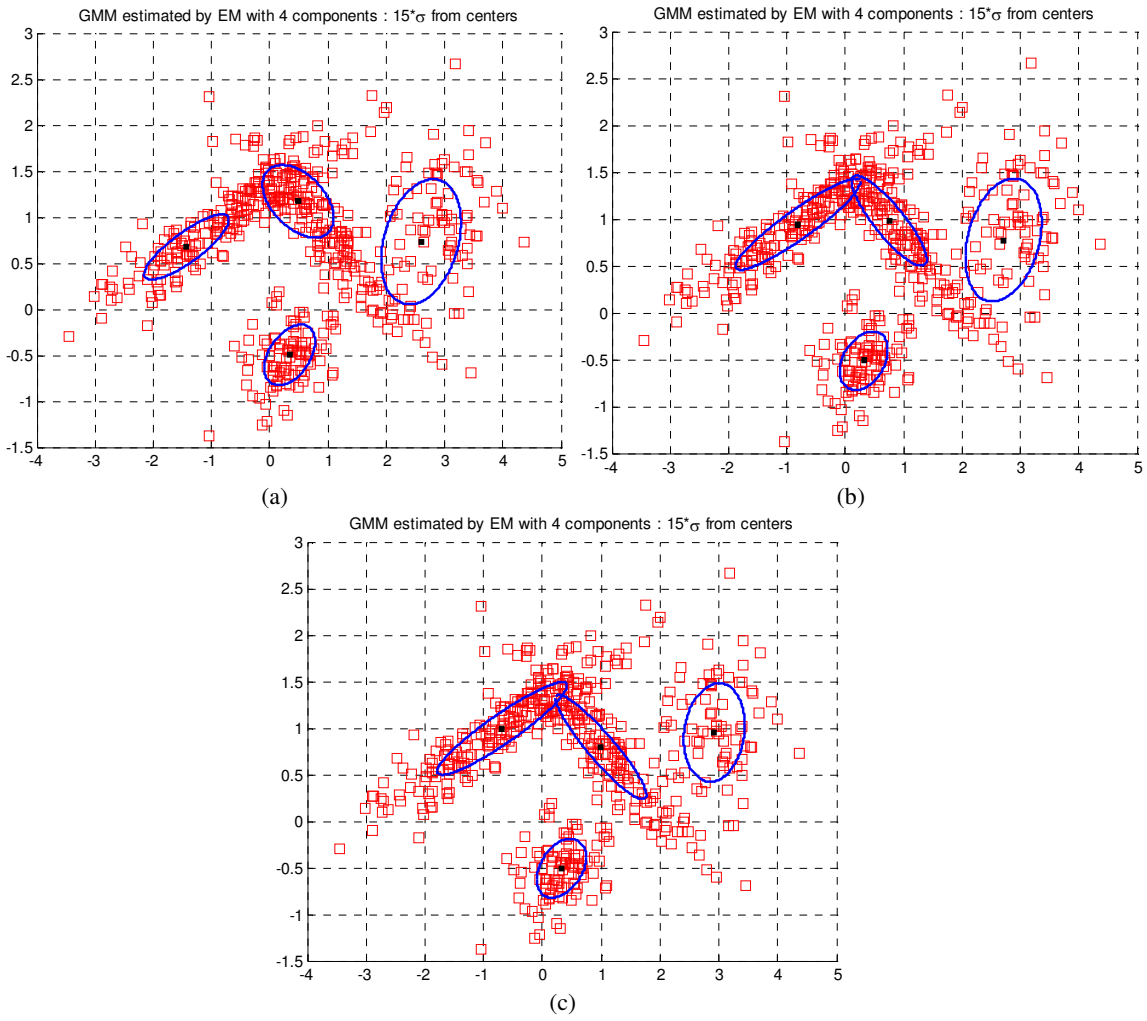
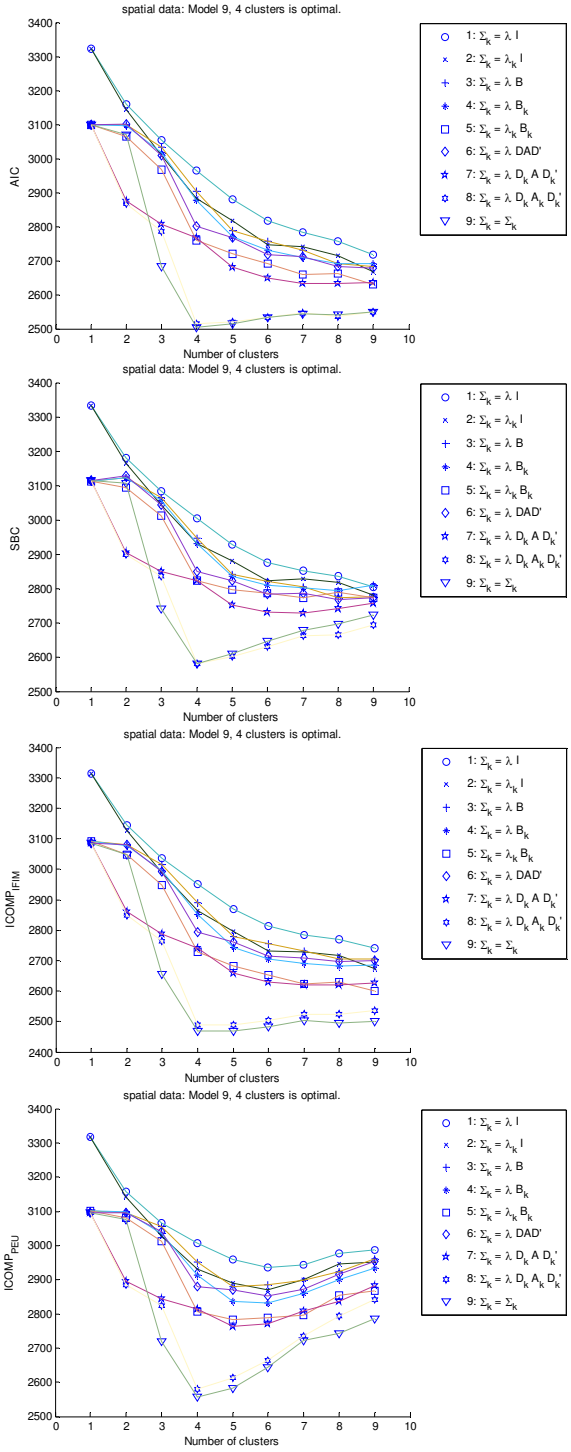


Figure 4.8: Estimated Gaussian components projected onto the data a) at iteration=1, b) iteration=6, c) iteration=13 for the synthetic data.



(a)

(b)

(c)

(d)

Figure 4.9: Model selection criteria scores a) AIC, b) SBC, c) $ICOMP_{IFIM}$, d) $ICOMP_{PEU}$ for the synthetic data.

In Table 4.3 model selection criteria scores for the unconstrained model (model [VVV]) and for a number of clusters $k = 1, \ldots, 9$ is given. All criteria agree on using 4 components.

Sample scatterplots of the data are given in Figure 4.10 with labeled estimated clusters. Also in Figure 4.11 surface plot of estimated Gaussian components is shown. The parameters related to the estimated Gaussians are given in Table 4.4.

Table 4.3: Model selection criteria scores from the best simulation for the unconstrained model for number of clusters $k = 1, \ldots, 9$ for the synthetic data.

Number of Clusters	AIC	SBC	$ICOMP_{IFIM}$	$ICOMP_{PEU}$
1	3097.90	3114.22	3085.94	3097.42
2	3070.81	3106.70	3046.79	3076.95
3	2684.17	2739.65	2655.44	2719.91
4	**2504.67**	**2579.73**	**2467.35**	**2557.87**
5	2515.52	2610.16	2468.25	2581.95
6	2532.19	2646.41	2484.77	2642.55
7	2542.77	2676.57	2503.19	2722.02
8	2541.60	2694.98	2495.02	2743.29
9	2549.47	2722.43	2499.62	2785.30

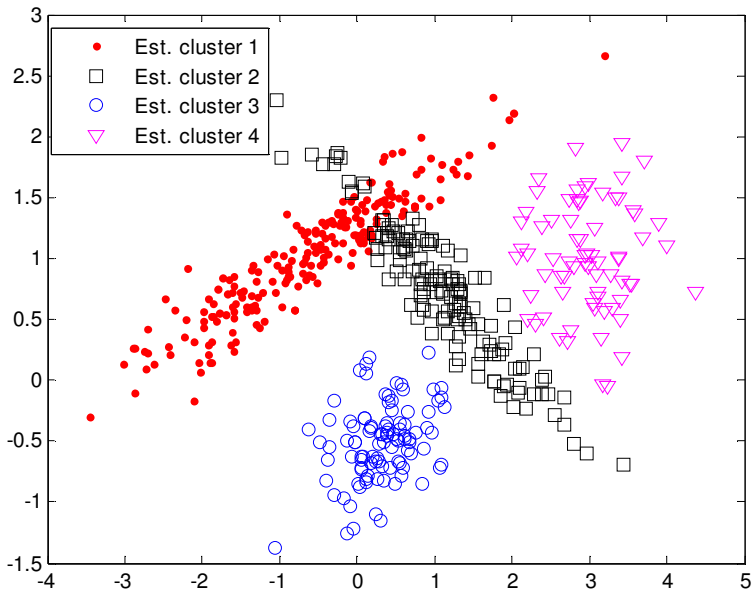


Figure 4.10: Scatterplot of the dataset with estimated labels of GMM based clustering for the synthetic data.

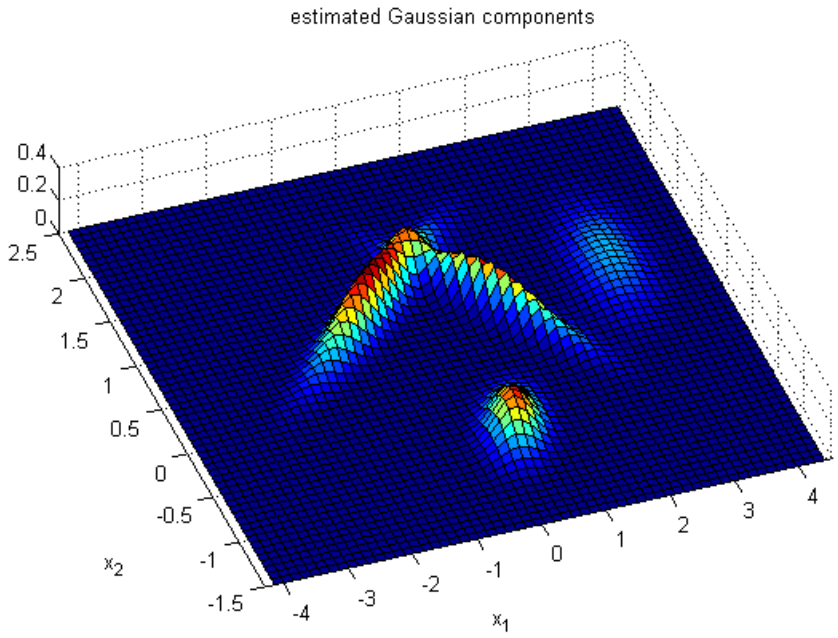


Figure 4.11: Surface plot of the estimated mixture density for the synthetic data

Table 4.4: Estimated Gaussian components' parameters for the synthetic data

Component	Approximated Mean Value, ($\hat{\mu}$)	Approximated Covariance Matrix ($\hat{\Sigma}$)	Approximated Weight ($\hat{w}$)
1	$[-0.6888; 0.9963]$	$\begin{bmatrix} 1.2184 & 0.5086 \\ 0.5086 & 0.2475 \end{bmatrix}$	0.3772
2	$[0.9884; 0.8043]$	$\begin{bmatrix} 0.5850 & -0.3935 \\ -0.3935 & 0.3109 \end{bmatrix}$	0.2820
3	$[0.3248; -0.5052];$	$\begin{bmatrix} 0.1742 & 0.0491 \\ 0.0491 & 0.0999 \end{bmatrix}$	0.1897
4	$[2.9154; 0.9441]$	$\begin{bmatrix} 0.2709 & 0.0433 \\ 0.0223 & 0.2908 \end{bmatrix}$	0.1512

To conclude the current experiment and as a passage to the next section, we can be asked to decide on which single point $(x, y)^T$ might be the best to represent the data consisting 525 points from 4 classes. Assuming the most dominant component of the mixture represents the main data and the others are all disturbance like information, μ value of the component having the highest weight can be favored to represent the data, which is $\hat{\mu}_1 = [-0.6888; 0.9963]$ as an estimation of the real value of $\mu_1 = [-0.7; 1]$.

Results of the Experimental Study of Background Modeling on Image Sequences

In our framework, after global motion compensation, our recent goal is the localization of moving objects. Segmenting out moving regions seen at each frame of the video is required, and then we can localize such regions. This is accomplished by examining the difference in pixel intensities between each new frame and

an estimation of the static background. Reliable background modeling which is critical for accurate identification of moving objects is more difficult when lighting conditions change. Here as the experimental study we will represent efficiency of GMMs based Background Modeling for general purpose video recordings. According to the literature GMM allows background modeling to evolve as the weather and time of the day affect lighting conditions. Steps for GMM based Background Modeling is given in Figure 4.12 which is a modified version of the flow diagram depicted in Figure 4.5.

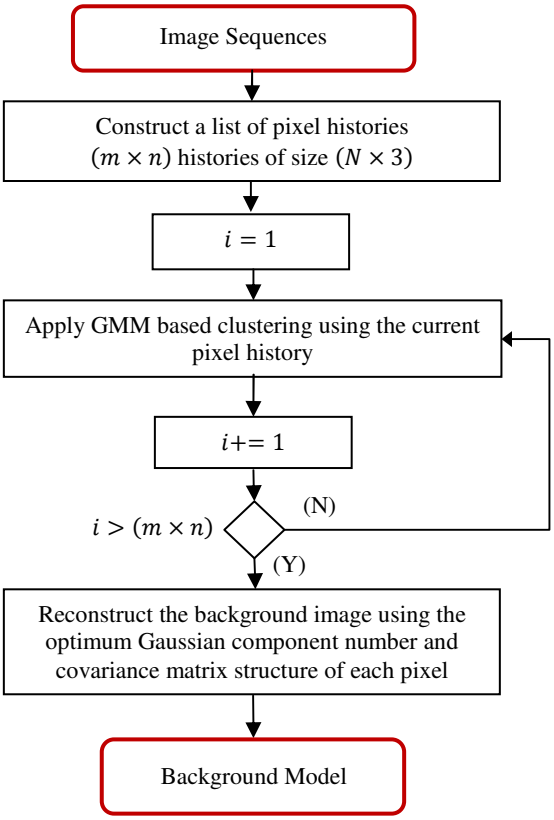


Figure 4.12: GMM based Background Modeling.

We want to continue our discussion on background modeling with the help of an illustration of just one GMM based pixel history clustering from "road surveillance video 2.1" pixel history at the location (99,235). Checking pixel histories for the data we decided not to use over 3 Gaussian components, in other words, we assume at each location pixel history can be described with at most 3 components. We repeated parameter initialization and EM for different numbers of $k = \{1,2,3\}$ and covariance forms and obtain the model selection criteria scores for each model. The model selection results for all four criteria scored, namely AIC, SBC, $\text{ICOMP}_{\text{IFIM}}$ or ICOMP and $\text{ICOMP}_{\text{PEU}}$ are given in Figure 4.14.

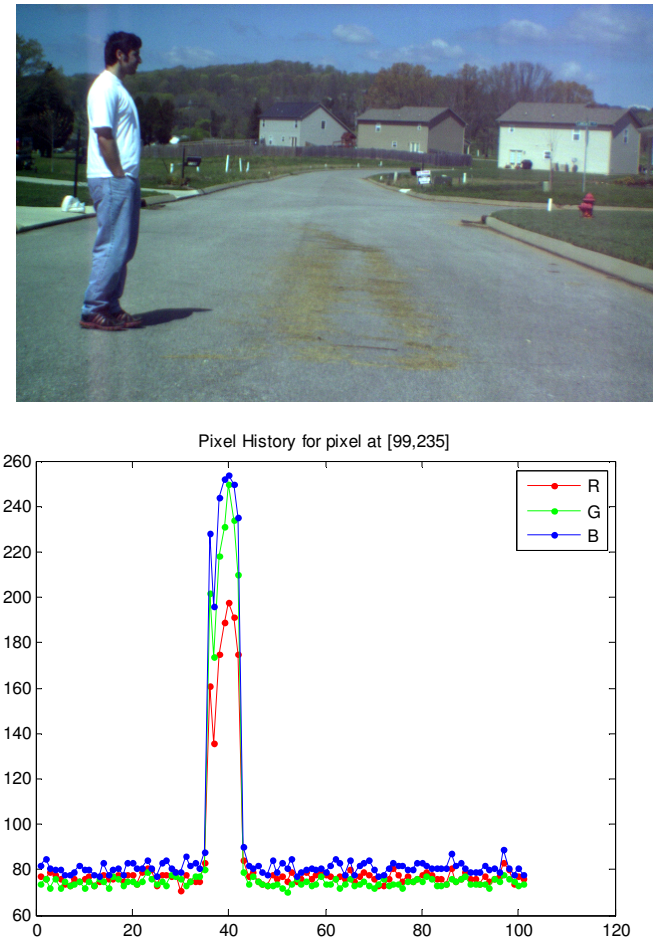
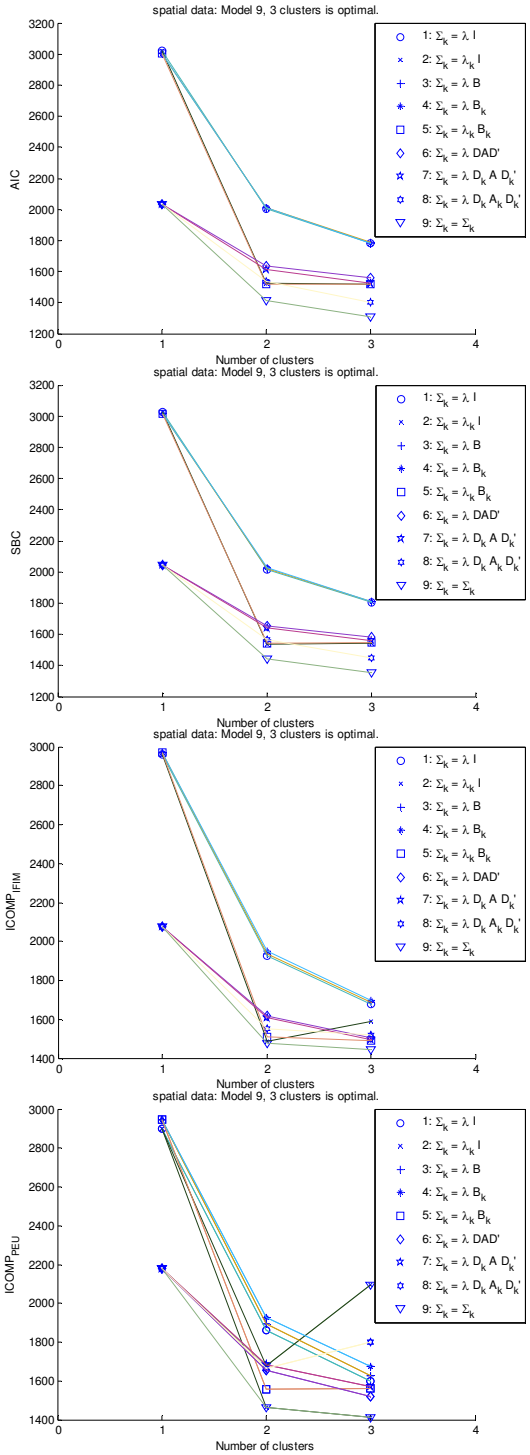


Figure 4.13: Test video example frames out of 101 frame-video "road surveillance video 2.1" 50% of each dimension (original 400x640)



(a)

(b)

(c)

(d)

Figure 4.14: Model selection criteria scores from the best simulation a) AIC, b) SBC, c) $ICOMP_{IFIM}$, d) $ICOMP_{PEU}$ for the pixel history data.

Table 4.5: Model selection criteria scores from the best simulation for the unconstrained model for number of clusters $k = 1, 2, 3$ for the pixel history data

Number of Clusters	AIC	SBC	$ICOMP_{IFIM}$	$ICOMP_{PEU}$
1	2031.55	2046.08	2076.23	2178.96
2	1411.30	1441.98	1476.97	1461.81
3	**1308.98**	**1355.82**	**1442.50**	**1412.52**

Table 4.6: Estimated Gaussian components' parameters for the pixel history data

Component	Approximated Mean Value, ($\hat{\mu}$)	Approximated Covariance Matrix ($\hat{\Sigma}$)	Approximated Weight ($\hat{w}$)
1	$\hat{\mu}_1 = \begin{bmatrix} 76.7664 \\ 74.6600 \\ 81.1707 \end{bmatrix}$	$\hat{\Sigma}_1 = \begin{bmatrix} 5.4133 & 1.8032 & 4.1461 \\ 1.8032 & 3.1607 & 1.5579 \\ 4.1461 & 1.5579 & 6.6305 \end{bmatrix}$	$\hat{w}_1 = 0.9304$
2	$\hat{\mu}_2 = \begin{bmatrix} 161.7505 \\ 201.0006 \\ 225.7505 \end{bmatrix}$	$\hat{\Sigma}_2 = \begin{bmatrix} 253.6800 & 259.7496 & 282.1789 \\ 259.7496 & 275.0093 & 299.7513 \\ 282.1789 & 299.7513 & 327.1791 \end{bmatrix}$	$\hat{w}_2 = 0.0397$
3	$\hat{\mu}_3 = \begin{bmatrix} 192.6671 \\ 238.3337 \\ 252.0004 \end{bmatrix}$	$\hat{\Sigma}_3 = \begin{bmatrix} 14.8878 & 32.1104 & 4.6677 \\ 32.1104 & 69.5576 & 10.6695 \\ 4.6677 & 10.6695 & 2.6674 \end{bmatrix}$	$\hat{w}_3 = 0.0299$

In Table 4.5 model selection criteria scores for the unconstrained model for number of clusters $k = 1, 2, 3$ is given. All criteria agree on using 3 components. Assuming the most dominant component of the mixture represents the background information of the current pixel, then the other components will be considered as the information related to moving objects occupying that pixel location. μ value of the component having the highest weight can be favored to represent background information at that location, which is $\mu_1 = [76.7664, 74.6600, 81.1707]^T$. The first component has a very high weight value $\hat{w}_1 = 0.9304$, there is no doubt that it dominates the pixels history data.

As stated in the flow diagram of Figure 4.12 repeating the process of GMM based clustering for every pixel we can compose a background image as a whole. Optimum K parameters for each pixels for "road surveillance video 2.1", "road surveillance video 2.2", "road surveillance video 2.3", and "road surveillance video 2.4" video data are shown in Figure 4.15. Background representations for the same data set are given in Figure 4.16.

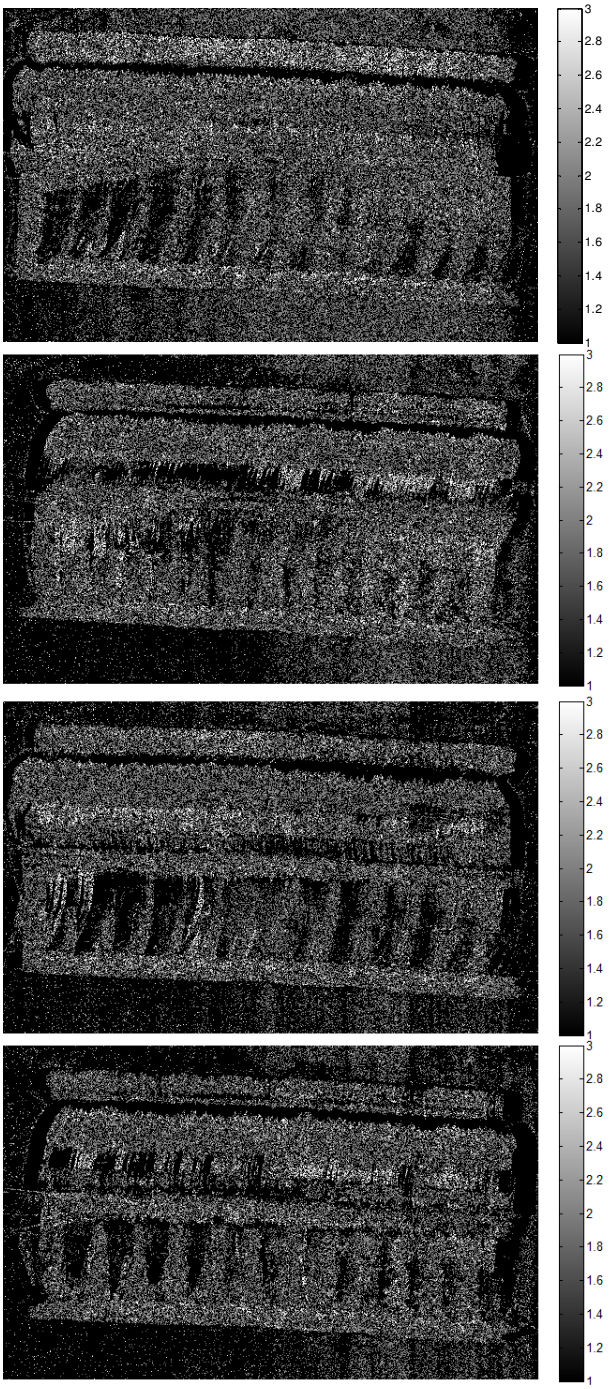


Figure 4.15: Optimum K values for every pixel's GMMs based Background Modeling for video dataset: road surveillance video 2.1, road surveillance video 2.2, road surveillance video 2.3, road surveillance video 2.4 45% of each dimension (original 400x640)

Figure 4.16: GMM based Background Modeling for video dataset: road surveillance video 2.1, road surveillance video 2.2, road surveillance video 2.3, road surveillance video 2.4 , 50% of each dimension (original 400x640)

As the second video dataset, we have “road surveillance video 1.1”, “road surveillance video 1.2”, “road surveillance video 1.3”, and “road surveillance video 1.4” in which the frames are originally affected by non-stationary camera system. Applying GMM based Background Modeling to the uncompensated image sequences gives us the background images given in Figure 4.16. As it can be seen clearly the background representations depend heavily on the assumption of the static background existence. In the case of non-stationary camera system global instabilities due to camera motion should be suppressed, and only the local displacements due to moving objects should be left out.

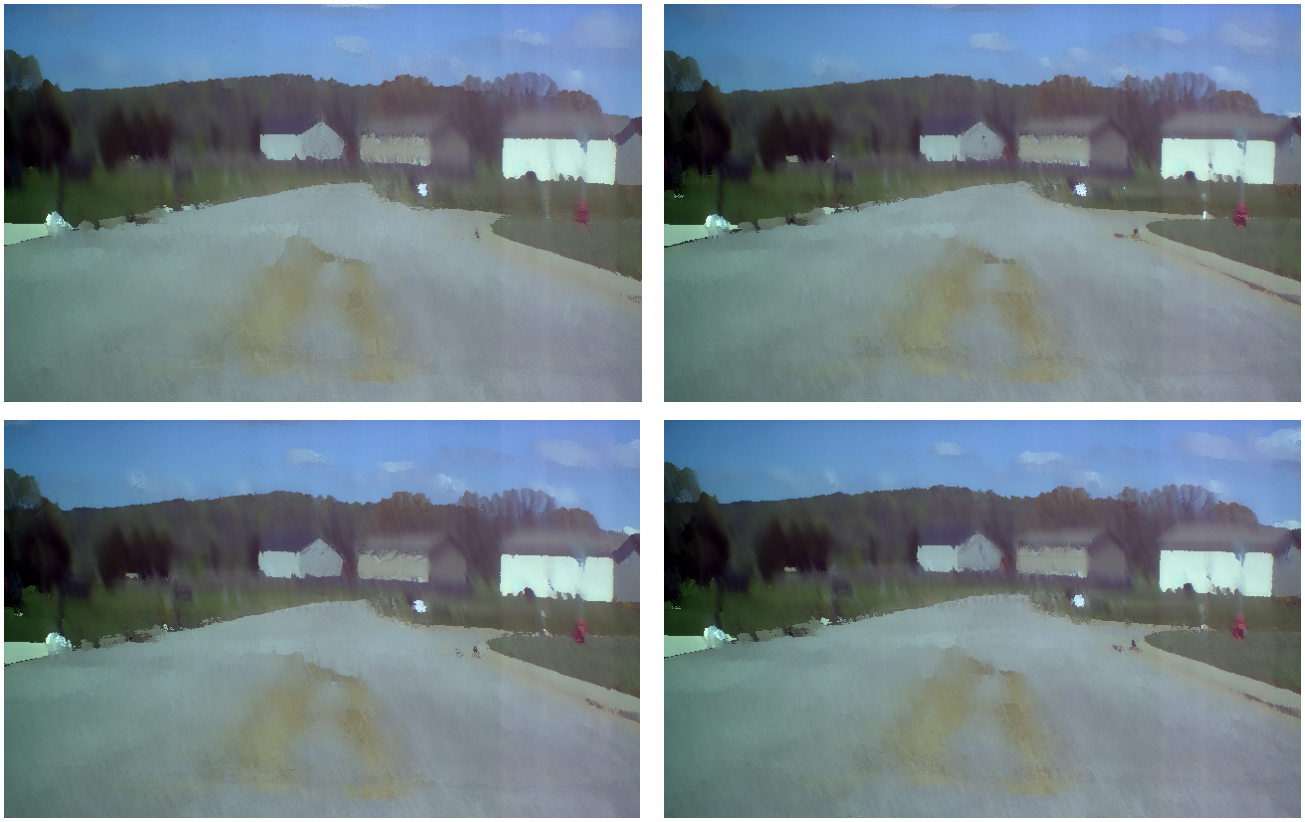


Figure 4.17: GMM based Background Modeling for the original video dataset: road surveillance video1.1, road surveillance video 1.2, road surveillance video 1.3, road surveillance video 1.4 , 50% of each dimension (original 400x640)

Global motion estimation process to stabilize the video should be carried out then GMM based background estimation can be used effectively. Optimum *K* parameter maps for each pixels for the stabilized "road surveillance video 2.1", "road surveillance video 2.2", "road surveillance video 2.3", and "road surveillance video 2.4" video dataset are shown in Figure 4.18. Background representations for the same dataset are given in Figure 4.18.

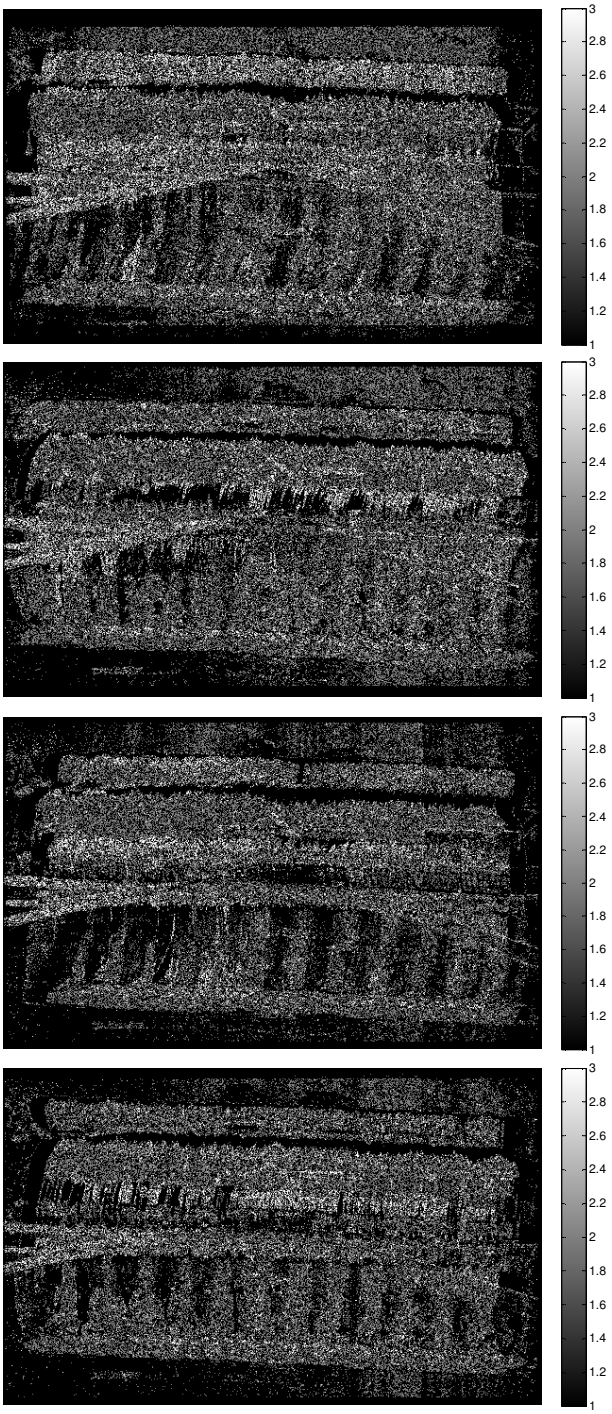


Figure 4.18: Optimum K values for every pixel's GMM based clustering for stabilized video data: road surveillance video 1.1, road surveillance video 1.2, road surveillance video 1.3, road surveillance video 1.4 50% of each dimension (original 400x640)

Figure 4.19: GMM based background estimation for the stabilized video data: road surveillance video1.1, road surveillance video 1.2, road surveillance video 1.3, road surveillance video 1.4, 50% of each dimension (original 400x640).

4.5 Summary

The main thrust of this dissertation research is Information Complexity guided GMMs for Statistical Background Modeling. The discussion held in the previous section as motion trajectories estimation aids to both background modeling which is detailed in this chapter to develop a motion segmentation scheme, and Super Resolution Image Reconstruction which will be discussed in the next chapter. We employed background subtraction which needs background modeling. We presented a new useful statistical technique; Information Complexity guided Gaussian Mixture Models (GMMs) based unsupervised Background modeling as the main contribution of our study, to extract moving objects in the frames. We introduced optimal K parameter as well as covariance model selection which produce the most dominant pixel values to compose a background model. This model is used for background subtraction in order to achieve motion segmentation which will aid Super Resolution Image Reconstruction. Background subtraction results will be presented in the next chapter. For each pixel we optimally identify the most important component, use the mean value of the component having the highest weight to represent the temporally changing pixel value at the current location. We achieved model parameter estimation using EM algorithm, which iteratively updates the parameters of the components which uses the K-means as an initialization step.

From optimal K maps for each dataset we can see that for the pixel locations highly occupied by the moving objects, regions higher number of K is needed.

5 Super Resolution Image/Video Reconstruction

In this study, Super Resolution (SR) Image/Video Reconstruction of the moving objects, is our ultimate goal. Contrary to traditional SR approaches, we employed several steps to compute the high resolution (HR) representations of the moving objects. We discussed suppression of the global motion trajectories imposed on the image sequence in Chapter 3, Gaussian Mixture Models (GMMs) based Background Modeling, which sets a base for motion segmentation accompanied by background subtraction, in Chapter 4. Background subtraction and moving object localization will be discussed briefly in the experimental study section of this chapter, yet the main discussion is super-resolving the accumulated information coming from multiple LR frames to reconstruct HR representations of the moving objects.

The straightforward idea of merely up-sampling and interpolating a single image does not produce a sufficient HR image. Single frame interpolation techniques have been researched quite extensively, with the nearest neighbor, bilinear, and various cubic spline interpolation methods providing progressively more accurate solutions. The goal of the most sophisticated members of this class is to magnify an image while maintaining the sharpness of the edges and the details in the image. In contrast in multi-frame Super SR, the goal is the recovery of missing high resolution that is not explicitly found in any individual low resolution (LR) image. Interpolation techniques to increase the size of a single image from an aliased LR image is inherently limited by the number of constraints available within the data and cannot recover the high frequency (specifically spatial frequencies) components lost or degraded during the LR sampling process. Despite the greater number of pixels after interpolation, the output image does not contain more details than the original observation. For this reason, single image interpolation methods are not considered to enhance the resolution. An intelligent approximation that enhances the high frequencies should be made. To achieve further improvements in this field, the next step requires the utilization of multiple datasets in which we use additional data constraints from several observations of the same scene. Temporarily correlated frames offer a better commencement than a single frame does towards improving the spatial resolution [Schultz96]. Here the resolution increase refers to up-sampling of the image thus increasing the maximum spatial frequency, undoing aliasing errors, and removing blur due to several effects.

A related problem to SR techniques is image restoration, which is a well-established area in image processing application and the literature on the restoration of a single input frame. In fact the two pathways are closely related, and SR techniques can be regarded as the second generation of image restoration, which takes advantage of using interdependency or temporal correlation of multiple frames towards adding abundant information to create HR images, as compared with that is available from a single image. Increasing the spatial resolution it as the heart of SR Image Reconstruction, opposed to the image restoration techniques. The techniques developed for single frame restoration have often provided the theoretical basis for extending to the SR techniques. Indeed, much of the work in single image restoration known as Single-Input-Single-Output (SISO) problem without resolution enhancement. In the SR area there are certain studies classified under the name single frame Super Resolution Image Reconstruction in which authors intend to use re-occurring patches within the single frame. In the theory there is not much difference between using multiple views of a region in multi-frames and using multiple views of the image

patches if they exist multiple times within a single image. In this study our motivation is always utilizing multi-frames. Multi-Input Single-Output (MISO) methods to estimate HR still image from LR observations used in the development of more general Multi-Input Multi-Output (MIMO) approaches to estimate HR image sequence. SR techniques may be applied to SISO problems as well as to the more general MISO or MIMO cases. The next level will be spatio temporal resolution enhancement of videos from LR frames which is not the direct focus of this study.

For static scenes the observations are related by global sub-pixel level displacements (due, for example, to the relative positions of the cameras, and to camera motion, such as panning or zooming), while for dynamic scenes certain regions of the scenes they are related by local sub-pixel level displacements due to object motion in addition to possibly global displacements. In both cases the objective of SR is again to utilize either the set of LR images to generate an image of increased spatial resolution [Katsaggelos07]. Moving objects gives a very important cue for human vision we can easily recognize objects as soon as they move. This kind of motion carries information about spatio-temporal relationships between objects in the field of a camera. For identifying objects that move or those entering or leaving the scene one also needs such information. We use the pixel differences aroused by motion as the cue for SR. Other cues also can be utilized to super-resolve a scene (for instance, observing the same scene with different blurs). The majority of the previous researches deal with some types of global displacement or rotation occurring between frames. This is rather impractical if a multi-frame technique is to be applied to an image sequence containing objects with independent motion trajectories. Several other topics are addressed in the previous chapters to accumulate all the information related to the extracted regions which are moving objects. In this chapter we will discuss how to use this information to obtain super-resolved representations using SR techniques.

Tsai and Huang [Tsai84] introduced the idea of employing SR. Having good representations of the scene despite the instabilities mentioned before by utilizing SR techniques has been intriguing the scientists since then. It could be said that the field of SR started in the sky with the launching of Landsat satellites [Katsaggelos07]. These satellites imaged the same region on the Earth and small displacements among the observations are approximated to provide a base for SR.

SR methodology is a well-posed problem since each LR observation from the neighboring frames potentially contains abundant knowledge about the desired HR image. Yet, it is unrealistic to assume that the super-resolved image can recover the original scene exactly. A reasonable goal of SR is a discrete representation of the original scene of that has a higher spatial resolution than the resolution of the available LR images. Estimated HR images are also expected to be free of possible degradations as well as the blurs due to the environment.

The main focus of this study is the real-world scenarios in which data from monitored scenes consist of the objects of interest is used. In this context reconstruction based methods fit conveniently to the problem domain. Reconstruction based methods rely on multiple LR images, and are the methods what come first to mind when SR is referred. Important elements of such techniques are the constraints imposed on the HR image representations of the scene through modeling of the observed LR images and the addition of prior information on the reconstruction.

Another class of SR techniques, which are recognition or learning based methods in the following sections, is out of focus of this dissertation.

5.1 Forward Image Acquisition Process and Scene Observation Models

It is necessary to first examine the forward process of image acquisition relating the desired HR image to LR observations or images. Cameras basically serve to record and preserve the scenes that are viewed through their lenses. Today's imaging systems record the time and space varying light intensity information rejected and emitted from objects in a three dimensional scene. Recording of an image sequence is achieved using an imaging system which is composed of an optical system, and a recording system [Borman02]. The optical system forms a two-dimensional image of the three-dimensional scene which reflects electro-magnetic radiation towards the camera. A series of lenses in the optical system focus the illumination on a two-dimensional surface, called the focal plane. Today's digital cameras provide the same function by recording images as digital information.

The idealized geometric properties of the ideal pinhole camera representation abstract away the complex process of optical image formation and replace it with purely geometric projection from locations in the 3-D scene to 2-D locations in the focal plane [Borman02]. Consider a point $\mathrm{X}(t) = [X(t), Y(t), Z(t)]^T$ in 3-D space. The optical system projects the 3-D point $X(t)$ onto the focal plane at position $\mathrm{x}(t) = [x(t), y(t)]^T$. The most commonly used model of the image projection characteristics is the perspective projection (Figure 5.1).

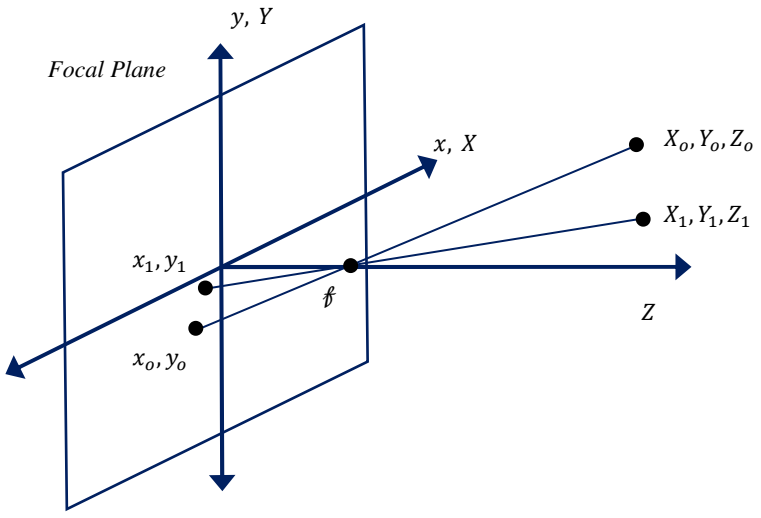


Figure 5.1: Perspective projection model for a pinhole camera with a focal length of f.

The electromagnetic radiation incident at the focal plane is a function of four continuous variables; two spatial variables, a temporal variable, and wavelength variable. A discussion from this point wavelength sampling is not the main interest of this study. The variation of the image as a function of time must be recorded. Typically we achieve this at regularly spaced time instants which are known as the temporal sampling. The sampling density in the temporal dimension is typically driven by the application and little can be done about it. Also, the variation of the light intensity is recorded at discrete locations. Digital image sequences of a video are consist of finite, regular 2-D lattice of picture elements which are the samples of the spatially varying illumination intensity pattern incident at the focal plane. Individual pixels may have one or more components to represent multi-spectral information. Also pixel values are quantized and stored using a finite bit-length digital representation.

Let us denote by the $f(x, y; t)$ the continuous (in time and space) dynamic scene projected onto the focal plane where $(\mathrm{x}; t) \in \mathbb{R}^3$ [Schultz94]. An HR image is the representation of $f(x, y; t)$ considering the sampling according to the Nyquist criterion in time and space. Thus reconstruction of the original signal utilizing appropriate reconstruction filters can be ensured. Real world scenes are usually not spatially band-limited (i.e. the Fourier transform (FT) of the signal has an unlimited support in frequency domain) low pass or so-called anti-aliasing filters are often used to ensure it is band-limited; in this way spatial sampling without artifacts is enabled. Area scan devices typically sample the entire image area over a single temporal integration period. As a result, there is a tendency away from interlaced capture and display systems. Consider HR image sensor plane, which is coincident with focal image plane, is divided into $qM \times qN$ square sensor elements, each of size $w/q \times w/q$. It is often not possible to impulse sample a function, that is, sample at a point. In reality sampling involves integration of the values in a spatio-temporal neighborhood. Each scan element outputs a discrete value which is proportional to the light which impinges upon a CCD sensor; each pixel accumulates the charge generated by photons which strike the light sensitive area of the pixel. For spatial sampling, this implies integration of the function over the spatial variables. Let $f[m, n]$ represents an HR image where $m = 0, \ldots qM - 1$ and $n = 0, \ldots qN - 1$ be the index of HR sensor measurements. These are computed from the continuous image via:

$$f[m,n] = \int_0^{q^2} \int_{\frac{w}{q}m}^{\frac{w}{q}(m+1)} \int_{\frac{w}{q}n}^{\frac{w}{q}(n+1)} f(x,y;t)dxdydt \tag{5.1}$$

Note that the measurements are accumulated for a short time $[0, q^2]$ which corresponds to temporal sampling or more specifically integration over time variable. Exposure time aperture of the camera lens determines the amount of incident illumination reaching to sensors. Integration over wavelength has side effects, since the sensing devices and materials respond to photons in a range of wavelengths rather than at discrete wavelengths. Let $g[i, j]$ represents a LR image where $i = 0, \ldots M - 1$ and $j = 0, \ldots N - 1$ be the index of LR sensor measurements. To keep the consistency we can assume this coarser grid consists sensor elements each of size $w \times w$. To have same level of intensities sensed by the sensors, acquisition time is changed from $[0, q^2]$ to $[0,1]$. This is computed from the continuous image via:

$$g[i,j] = \int_0^1 \int_{wj}^{w(j+1)} \int_{wi}^{w(i+1)} f(x,y;t)dxdydt \tag{5.2}$$

The direct relationship between HR and LR images (Figure 5.2) can be formulated as;

$$g[i,j] = \frac{1}{q^2} \sum_{m=qi}^{q(i+1)-1} \sum_{n=qj}^{q(j+1)-1} f[m,n] \tag{5.3}$$

Consider g as $MN \times 1$ lexicographically ordered vector that contains pixel values from the LR image and f as the $q^2MN \times 1$ vector containing pixel values from the HR image. The decimation system model in Equation 5.3 can be written in vector-matrix form as $\mathrm{g} = \mathrm{Df}$, where D is the $MN \times q^2MN$ size decimation

matrix. If there is inherent noise affecting the imaging system then system should be analyzed considering the noise vector n as:

$$g = Df + n \tag{5.4}$$

The scene can be represented in time ideally according to the Nyquist criterion as a sequence of HR images, $f_k,\ k = 0, \dots, K-1$. Due to physical limitations what we get out of the imaging system mostly is a LR sequence $g_k,\ k = 0, \dots, K-1$. The vast majority of the SR algorithms use a short sequence of LR input frames to produce a single super-resolved high-resolution output frame (the MISO case). The objective of SR addressed in this dissertation is to obtain an estimate of one HR frame f_k at each time from available observations. The same techniques may, however, be applied to resolution enhancement of videos by using a shifting window of processed LR frames utilizing sliding window approach as illustrated in Figure 5.3. Sliding window determines the subset of low-resolution frames to be processed. The window is moved forward in time to produce successive super-resolved frames in the output sequence. Various approaches may be taken to determine the subset of low-resolution frames used to compute the HR frames corresponding to the start and end of the observed low-resolution image sequence. A contradiction to this approach is employing sequential SR methodology in which previously estimated HR frames as well as a number of LR images are used together.

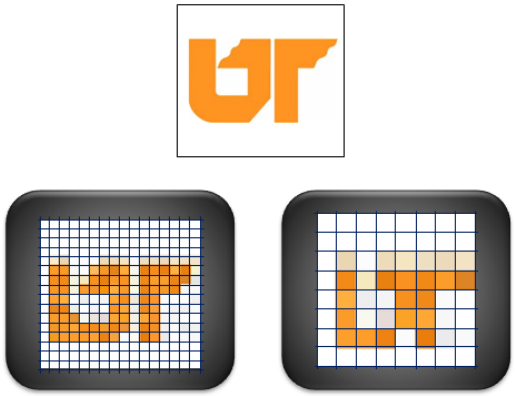


Figure 5.2: Relative HR and LR representations of the practically continuous scene related to CCD sensor sizes.

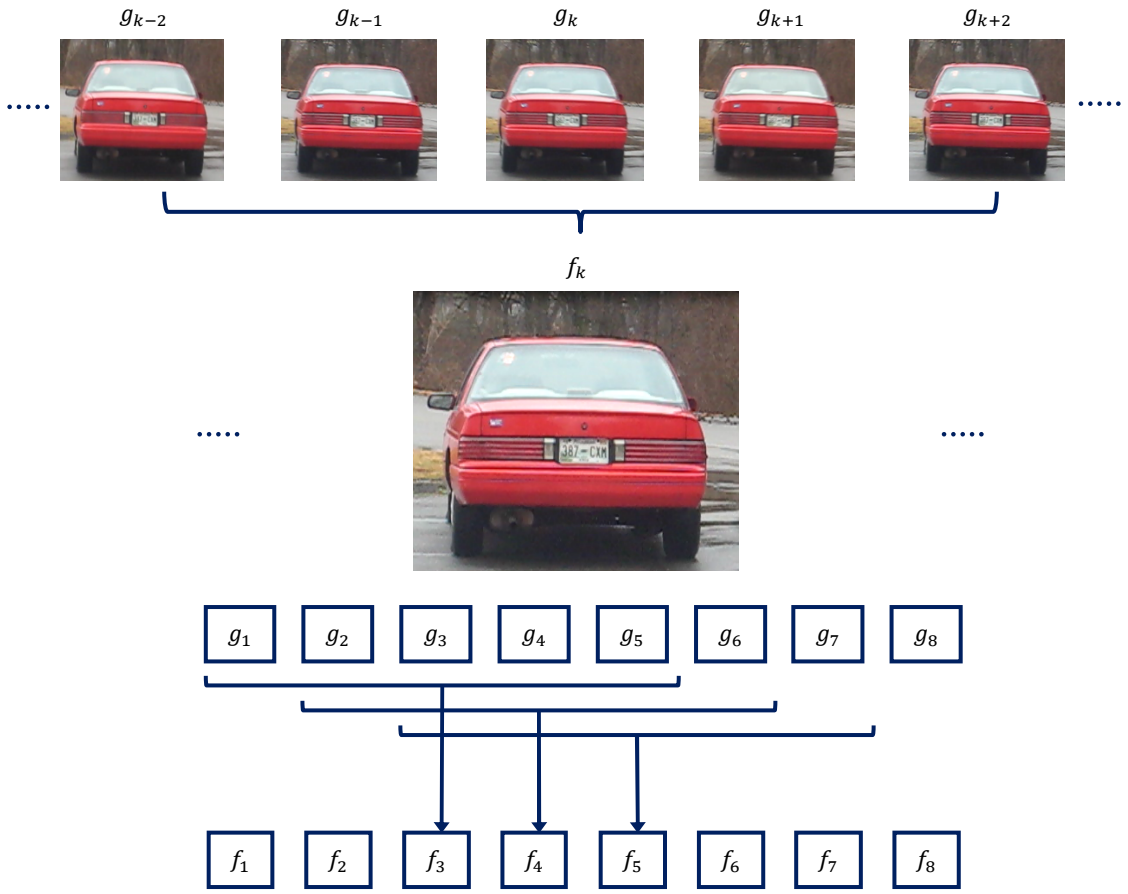


Figure 5.3: Sliding window approach; using a number of LR observations to estimate an HR representation of the scene.

5.2 Image Acquisition Scenarios

There are several issues need to be addressed in a practical situation to complete the observation model in Equation 5.4 given as the simplest scenario and an ideal situation. Due to the spatial sampling rate quality of the image acquisition devices and instability of the observed scene (movement of local objects, vibrating imaging systems, media turbulence, change of focus, and motion blur due to low shutter speed); the acquired images suffers from aliasing, blurring, presence of noise and insufficient spatial resolution. In applications such as astronomy, medicine or physics one is faced with images which the noise reduction is the main issue. The focal plane image is geometrically deformed or warped when generating the frames. Several optical system problems such as out of focus blur and relative camera-scene motion blur effect images again. The latter effects are commonly modeled via convolution (or linear shift invariant –LSI- filtering) of the image with an unknown point spread function (PSF). An interesting modeling question is the order in which these two operations –blurring and warping- are applied. Both systems, so-called warp–blur and blur–warp models are given in the following discussions. Finally, the CCD discretizes the images and produces digitized noisy image or frame. The aliasing effect will be present in the LR images of the original HR image after decimation as long as the high frequencies are not cut by the system. The noise

component can model various elements in the imaging chain, such as, the thermal or electronic noise, or the errors during storage or transmission. CCD scan acts as convolution followed by sampling operator $\mathcal{S}$, $\mathcal{D}(\cdot) = \mathcal{S}(h \otimes \cdot)$ In conclusion, the observation model for k^{th} LR frame up to an SIMO observation model is formulated in [Sroubek07] as:

$$g_k[m,n] = \mathcal{D}\left(v_k(x,y) \otimes f\left(\mathcal{W}_k(x,y)\right)\right) + n_k[m,n] \tag{5.5}$$

where $\mathcal{D}$ is the decimation operator that models the function of the CCD sensors and consists of convolution process with sensor's PSF, h. It is a nonlinear function which digitizes and decimates the function into pixels values from continuous intensities into a number of gray levels. $\mathcal{W}_k$ is the complex geometric deformation or warping, f is the focal plane image v_k is representing the optical system blurs, n_k is the additive noise, $[m,n]$ is the image grid, (x,y) is the world coordinates, and the index k represents the discrete temporal instants. The original continuous focal plane image f is a single input and the acquired discrete LR images g_k $(k = 1 \dots frN)$ are the multiple outputs. This is a very realistic yet not the most useful formulation in SR methodology. They also adopted the idea of the blurring to be space invariant for the sake of simplicity, we can add one or assumption that through multiple observations, decimation operator remains the same.

Possible HR representations of the scene, which are assumed to be sampled greater than or equal to Nyquist rate from a continuous focal plane image, should be estimated using the available LR image sequence. According to the reasonable goal of SR, source of the LR observations can be regarded as the discrete space HR image. This idea is formulated next to get rid of the ambiguity on what SR techniques promise. The continuous focal plane image f of Equation 5.5 is replaced for lexicographically ordered vector f, the HR representation of scene. Other vector-matrix formulations of the quantities in Equation 5.5 [Papathanassiou05, Baker02] will be as following:

$$\mathrm{g}_k = \mathrm{DV}_k\mathrm{W}_k\mathrm{f} + \mathrm{n}_k. \tag{5.6}$$

The f vector of size $q^2MN \times 1$, where q is the down-sampling factor. f is the lexicographical discrete representation of the continuous focal plane image f and is assumed to be subjected to the same series of degradations which are represented in vector-matrix formations. W_k is the warping matrix, V_k is optical system degradations matrix, and D is the decimation matrix to generate aliased LR frames g_k of size $MN \times 1$. N_k denotes a noise field. Noise elevates on each pixel of the $M \times N$ size observed image along temporal space. It is a set of random variables.

The formulation in Equation 5.5 or Equation 5.6 is more applicable to the scenario where several cameras acquire still images of the same scene which are then combined to produce an HR image. A generalized version of this formulation is also used commonly [Leung08, Zibetti05] under certain assumptions:

$$\mathrm{g}_k = \mathrm{A}_k\mathrm{f} + \mathrm{n}_k \tag{5.7}$$

where A_k matrix of size $MN \times q^2MN$ represents the behavior of the system for k^{th} LR observation. It contains blurring, warping, and down-sampling processes altogether. According to the temporally non-coincident observation model [Borman99] we can generalize Equation 5.6 as:

$$\mathrm{g}_l = \mathrm{DV}_l\mathrm{W}_{l,k}\mathrm{f}_k + \mathrm{n}_{l,k}. \tag{5.8}$$

This model suggests the warping of an image is applied before it is blurred. The end-to-end system in this case is depicted in Figure 5.4. Another acquisition model used in the literature first considers the blurring of the HR representation of the scene followed by warping and down-sampling or decimation operation as shown in Figure 5.5. In this case the observation model becomes:

$$\mathrm{g}_l = \mathrm{DM}_{l,k}\mathrm{B}_k\mathrm{f}_k + \mathrm{n}_{l,k}. \tag{5.9}$$

where $n_{l,k}$ denotes the acquisition and registration noise, B_k the blurring matrix for the k^{th} HR image, $M_{l,k}$ the motion compensation operator for the blurred HR images and D again is the down-sampling or decimation matrix. Different notation has been used in Equation 5.8 and Equation 5.9 for the blur and warping operators in order to distinguish these two models. The question as to which of the two models (blur–warp or warp–blur) should be used is addressed in [Wang04]. The authors claim that when the motion has to be estimated from the LR images, using the warp–blur model may cause systematic errors and, in this case, it is more appropriate to use the blur–warp model. They show that when the imaging blur is spatio-temporally shift invariant and the motion has only a global translational component the two models coincide.

In particular, the performance of SR methods depends on a complex relationship between the measurement signal to noise ratio (SNR), the number of observed frames, the set of relative motions between frames, the image content, and the PSF of the system.

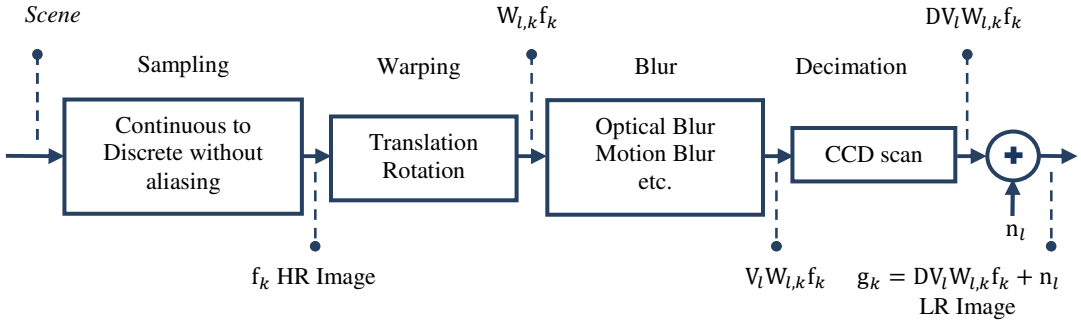


Figure 5.4: Temporally non-coincident warp-blur observation model.

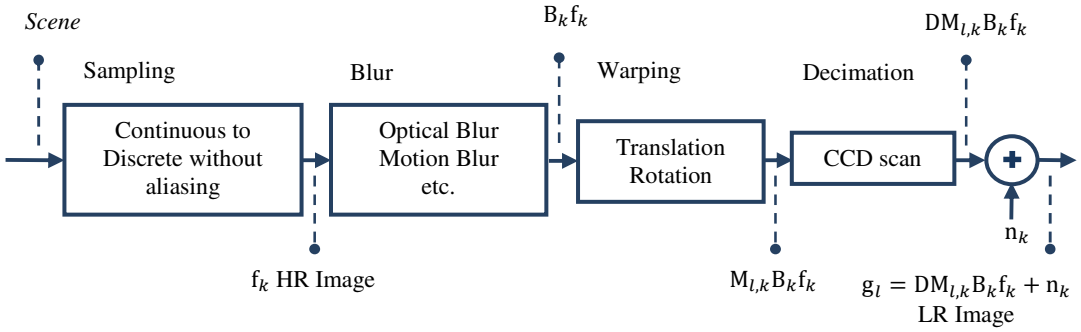


Figure 5.5: Temporally non-coincident blur-warp observation model.

5.3 SR Methodology Used in This Study

To obtain super-resolved HR image from LR observations, the acquisition model in Figure 5.4 which covers three distinct cases, are studied differently in literature [Sroubek07]. If we want to resolve the geometric degradation, we face a registration problem. Second, if the decimation operator D and the geometric transform $W_{l,k}$ are not considered, we face a multi-frame blind deconvolution problem. Third, if the optical system blurs are not considered or assumed known, and $d_{l,k}$ parameters are suppressed up to a sub-pixel translation, we obtain a classical SR formulation. In practice, it is crucial to consider all three cases at once which is difficult to handle. We deal with warping and decimation in this study; sub-pixel motion compensation to un-warp, and multi-frame image interpolation/reconstruction to un-decimate the LR images towards SR Image Reconstruction. As mentioned before, segmented out and localized moving objects from the image sequences are referring to LR images used in the Multi-frame Image Reconstruction stage, in this study.

We adopted the SR Methodology of Vandewalle et al. [Vandewalle06]. They state that there are two major independent challenges of SR. LR images differ from each other by local and global planar motions. Therefore, the first challenge corresponds to having precise knowledge of motion parameters, an assumption which does not favor PSF usage, and suggests only motion suppression. In real-life imaging applications, the motion occurring between frames is not known exactly, since precise control over the data acquisition process is rarely available. Thus, motion estimations must be computed to determine sub-pixel displacements between frames. The quality of these motion estimates will have a direct effect on the quality of the enhancement algorithm. The artifacts caused by incorrectly aligned LR frame set are visually more disturbing than the degradation seen only interpolating a single image. If enough frames with the correct sub-pixel displacements are available, then the second challenge multi-image interpolation problem is no longer ill posed. In other words, a unique solution can be obtained.

Standard SR approaches consist of two stages and this is what we utilize in this study as SR Image Reconstruction; first the LR images are aligned onto the same coordinate system through sub-pixel registration. The important assumption is that no occlusion is present if the depth variation on the scene is planar. After estimating motion differences the information obtained from multiple images are used to the reconstruction of a sharp HR image. Interpolation onto a uniform grid is done to obtain a uniformly spaced up-sampled image. LR images are overlaid on an HR grid, and missing values are wisely interpolated (Figure 5.6). These stages can be implemented separately or simultaneously according to the reconstruction method adopted. We utilized these steps separately.

A critical component in the system modeling the generation of the LR observations from HR source data is the warping system. There are many warp models used in the SR literature as well as many techniques for the estimation of their parameters. In the context of dynamic video scenes the difference between frames is most probably due to global motion of the camera and locally moving objects in the scene. Almost all of the SR methods proposed in the literature use the slight motion estimation as the most fundamental cue for estimating the HR images. Given a sequence of images they are registered with sub-pixel accuracy in respect to translation and rotation. This opens a pathway for image enhancement in respect to improved resolution. However the assumption of the use of slightly different low-resolution images of the same scene to construct a higher resolution image is not always practicable. This slight difference is mostly assumed to have some common origins: camera vibration, change of focus, or a combination of these. In this study, the level of displacements we deal with is too high compared to the scenarios studied in the state of art methods. An illustration of the level of displacements is shown with the help of edge images superimposed on the reference frame in Figure 5.7.

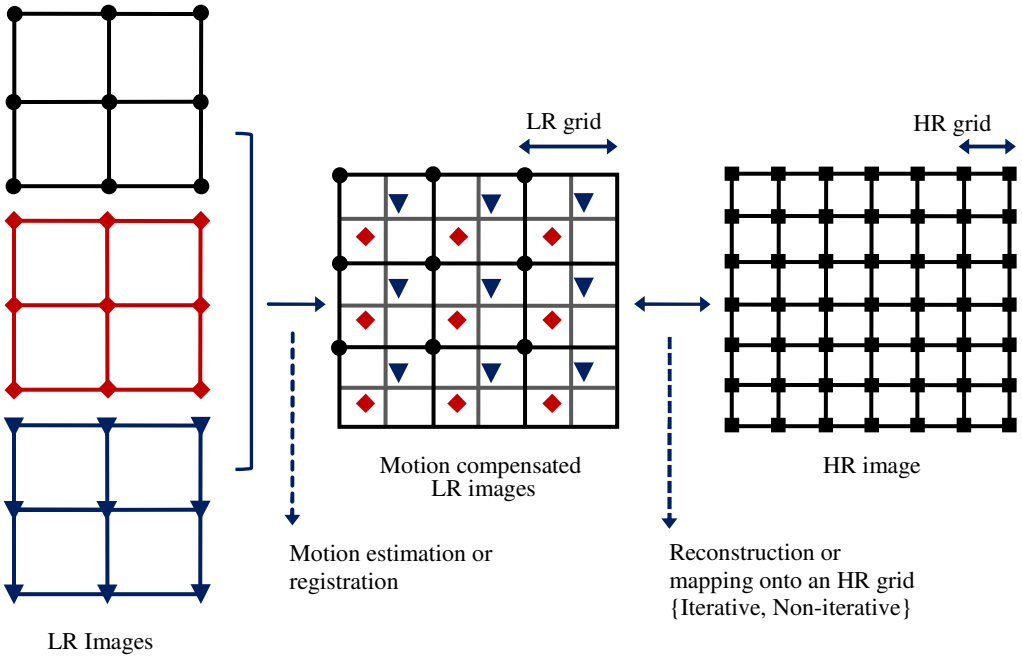


Figure 5.6: Basic premise for traditional SR methods; all frames are aligned onto the reference frame – top left-. Sub-pixel registration takes place. Registered LR images are used for SR Image Reconstruction.

a) b)

Figure 5.7: a) Level of displacements using background subtracted LR edge images of 4 frames of road surveillance video 2.1. 50% of each dimension (original 400x640), b) Classic SR Image Reconstruction result, (cropped from 800x1240)

5.4 Statistical and Spatial Analysis of the Aligned Information

Accurately registered LR images can be combined to reconstruct an HR representations in the reconstruction step depicted in Figure 5.6. We employed previously proposed state of art SR Image Reconstruction methods, with a focus on the Kriging method. A list of the SR Image Reconstruction methods that we are using is given as follows; readers can use the references for the comprehensive understanding of each method.

- Interpolation: This method simply locates all the images' pixels on an HR grid using a fitting function.
- Papoulis-Gerchberg [Papoulis77]: Papoulis and Gerchberg's algorithm is projecting the accumulated information successively onto the space of known pixels and the space of band-limited images.
- Iterated Back Projection [Keren88]: The idea behind Iterated Back Projection is to start with a rough estimation of the HR image, and iteratively add to it a "gradient" image, The sum of the errors between each LR image and the estimated HR image that went through the appropriate transforms is used.
- Robust Super Resolution [Zomet01]: Robust Super Resolution is a more robust version of the Iterated Back Projection. The only difference resides in the computation of the gradient, which is not given by the sum of all errors, but by the median of all errors. This brings robustness against outliners in the LR images.
- POCS [Patti97]: Projection onto Convex Sets (POCS) algorithm defines convex sets expressing constraints on the reconstructed image. Estimated reconstruction is successively projected onto different convex sets.
- Structure-Adaptive Normalized Convolution [Pham2006]: It is a framework that combines a maximum likelihood/maximum a posteriori (MAP) approach with a POCS approach to define a new convex optimization process.

Spatial Analysis Employing Kriging

The need for spatial analysis and Kriging will be clarified in this section. First we want to mention about several concepts and statistical tools. In Kriging we employ variograms. The variogram characterizes the spatial continuity or roughness of a dataset.

Ordinary one dimensional statistics for two datasets may be nearly identical, but the spatial continuity may be quite different. Some common descriptive statistics for the datasets are number of samples, sample mean, sample median, sample covariance matrix, standard deviations etc. It is not hard to show two datasets showing exactly the same descriptive statics yet so different from each other (Figure 5.8). However, these two datasets are significantly different in ways that are not captured by the common descriptive statistics and histograms. The visually apparent difference between these two datasets is due to one of texture and not variability.

Variogram analysis consists of the experimental variogram calculated from the data and the variogram model fitted to the data [Barnes11]. The experimental variogram is calculated by averaging one half the difference squared of the z-values over all pairs of observations with the specified separation distance and direction. It is mostly plotted as a 2-D graph. Consider a scatterplot where the data pairs represent measurements of the same variable made some distance apart from each other. The separation distance is usually referred to as "lag", as used in time series analysis. Unlike the researches on time series, in which either the covariance function or the correlogram is highly used, the researches on spatial analysis utilizes typically the semi-variograms. This is primarily because the semi-variogram, which averages squared differences of the variable, tends to filter the influence of a spatially varying mean.

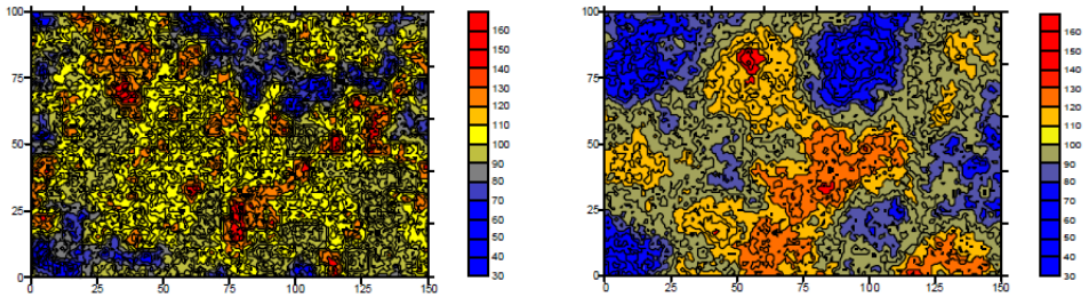


Figure 5.8: Datasets having different texture regardless of their identical statistical descriptors [Barnes11].

The variogram model is chosen from a set of mathematical functions that describe spatial relationships. The appropriate model is chosen by matching the shape of the curve of the experimental variogram to the shape of the curve of the mathematical function. The geometric anisotropy of the data can be accounted for (variable spatial continuity in different directions). Separate experimental and model variograms can be calculated for different directions in the dataset. Actually, since we are mainly dealing with digital image sequences, there is no type of anisotropy present.

We formally used geostatistical inference methods in under the name Kriging [Krige51] at the last stage of our current framework Motion Segmentation aided SR Image Reconstruction to super–resolve the LR images. The advantages of Kriging are twofold; it provides estimates of the values at unknown locations with a minimum error and it is a completely data-driven approach. Kriging defines a stochastic process model, under which a wise-interpolation is done.

Kriging has been proven to outperform all other interpolation methods – under specific conditions (e.g., when the relationship between the data can be readily modeled by a parametric function) – and not to perform worse. The overall process for Kriging consists of three steps: estimating the spatial correlation between the measured samples, constructing an ideal model that best fits the estimated spatial correlation, and estimation of the new values using Kriging. In the simple case of regularly sampled data, the computation of the semi-variance is quite straightforward. Assuming the sampling interval (lag) is d, the semi-variance for distances equal to multiples of d can be computed as [Grinstead07]:

(5.10)

where is the measurement of a regionalized variable taken at location , is another measurement taken intervals away, is the number of points used per lag interval.

Once the experimental variogram has been calculated, an ideal parametric model is fit to the data through an automatic optimization method. Least-squares fitting of a number of ideal models is performed, and the one with the best match to the data is used as the ideal variogram model for the Kriging process. Kriging is the actual process of using the parametric variogram model to estimate the value at the specified location. The most common form of Kriging used in engineering applications is punctual (point) Kriging – where the estimate for a single point is calculated from the values of nearby points. In punctual Kriging, the estimate of an unknown value uses a weighted summation of other nearby known points:

$$Z_e(p) = \Sigma\, w_i Z(p_i) \tag{5.11}$$

The error associated with this estimate Z_e and the actual value Z_a at this location is. $\varepsilon_p = Z_e(p) - Z_a(p)$. Ideally, Kriging attempts to minimize this error. The variance of this error is the amount of scattering of the estimates about their true values:

$$\sigma_z^2 = \frac{\sum_{i=1}^{n}\left(Z_e(p_i) - Z_a(p_i)\right)^2}{n} \tag{5.12}$$

The estimation and its error are dependent on the weights chosen in Equation 5.11. Optimal weights, therefore, would be those that produce the minimum estimation variance. These are found by solving a system of equations consisting of the weighted semi-variances between measured points, and the estimated semi-variances between the unknown values and the known values:

$$\begin{bmatrix} \gamma(d_{11}) & \dots & \gamma(d_{1n}) \\ \vdots & \ddots & \vdots \\ \gamma(d_{n1}) & \dots & \gamma(d_{nn}) \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} \gamma(d_{1p}) \\ \vdots \\ \gamma(d_{np}) \end{bmatrix} \tag{5.13}$$

As a simple example, let us estimate an unknown value $Z_e(p)$ using the known values Z_1, Z_2, Z_3, and, Z_4. Since we have 4 points that will contribute to the estimation, 4 weights must be determined. Thus, we have 4 simultaneous equations:

$$\begin{aligned} w_1\gamma(d_{11}) + w_2\gamma(d_{12}) + w_3\gamma(d_{13}) + w_4\gamma(d_{14}) &= \gamma(d_{1p}) \\ w_1\gamma(d_{21}) + w_2\gamma(d_{22}) + w_3\gamma(d_{23}) + w_4\gamma(d_{24}) &= \gamma(d_{2p}) \\ w_1\gamma(d_{31}) + w_2\gamma(d_{32}) + w_3\gamma(d_{33}) + w_4\gamma(d_{34}) &= \gamma(d_{3p}) \\ w_1\gamma(d_{41}) + w_2\gamma(d_{42}) + w_3\gamma(d_{43}) + w_4\gamma(d_{44}) &= \gamma(d_{4p}) \end{aligned} \tag{5.14}$$

where $\gamma(d_{ij})$ is the semi-variance between points i and j, and d_{ij} is the distance between the two points. The semi-variance values are taken from the parameterized variogram. To assure that the solution is unbiased, a further constraint of $\sum w_i = 1$ is usually applied. This leads to an over-constrained system, so another variable is added to the system, called the Lagrangian multiplier, to insure a minimum error solution is obtained. The weights that are the solution of this system are then plugged into Equation 5.11 to estimate the value for the point of interest [Grinstead07]. Thus, in the general form, the Kriging equations are:

$$\begin{bmatrix} \Gamma_z & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} w \\ \lambda \end{bmatrix} = \begin{bmatrix} c_z \\ 1 \end{bmatrix} \tag{5.15}$$

where Γ_z is the semi-variance matrix taken from the semi-variogram, c_z is a vector of the observed values used for Kriging, w is the solution for the weights of the ordinary Kriging estimator, and λ is the Lagrange multiplier. During the Kriging process the only user-specified parameters are the sampling interval (lag) for measurement and the library of variogram functions provided to work from. The experimental semi-variances can be fit to the "best match" variogram in the library. Also, Kriging's accuracy can be improved with prior knowledge of the system and the addition of some user-specified constraints.

5.5 Experimental Results

We deal with dynamic scenes with both stationary and non-stationary camera systems. As stated earlier in the introduction chapter, our main assumption related to the scene is; for stationary camera system moving regions of the dynamic scene due to object motions, are related by local displacements, for non-stationary

camera systems these regions are related by local displacements, in addition to possible global displacements can be imposed on the whole scene.

Using these multiple frames, the unnecessary information coming from background is aimed to be eliminated by utilizing motion segmentation. Removing the restriction of regularly shaped regions leads to region-based motion models to localize segmented our moving objects. The partitions can be arbitrary shapes obtained as the outputs of the segmentation algorithm. This kind of structure is allowed to adopt over time to track the apparent motion. Region-based motion representations often provide accurate and efficient motion representations. Then we compute rough local motion estimation of moving objects, which are already separated from background. After finding the correspondences in the LR images, we force the corresponding scene pixels of multi-frames to be tightly close to each other. As the final step next discussions will be about SR Image Reconstruction of the set of moving objects, which are ensured to come tightly close to each other.

In the previous chapter as the experimental study we represented efficiency of Information Complexity guided Gaussian Mixture Models (GMMs) for Statistical Background Modeling for general purpose video recordings, which we have global motion present related the camera itself. In real applications due to relative motion of the moving object, a pose change is inevitable. Using sequential frames we are aiming not to be trapped into such a problem. The non-rigid regions in the scene are basically not covered, and in principle they are disregarded. The motions of the lower body show obviously a non-rigid behavior. For several video datasets the frames related to a moving object at 9 different time instants are given in Figure 5.9-16.

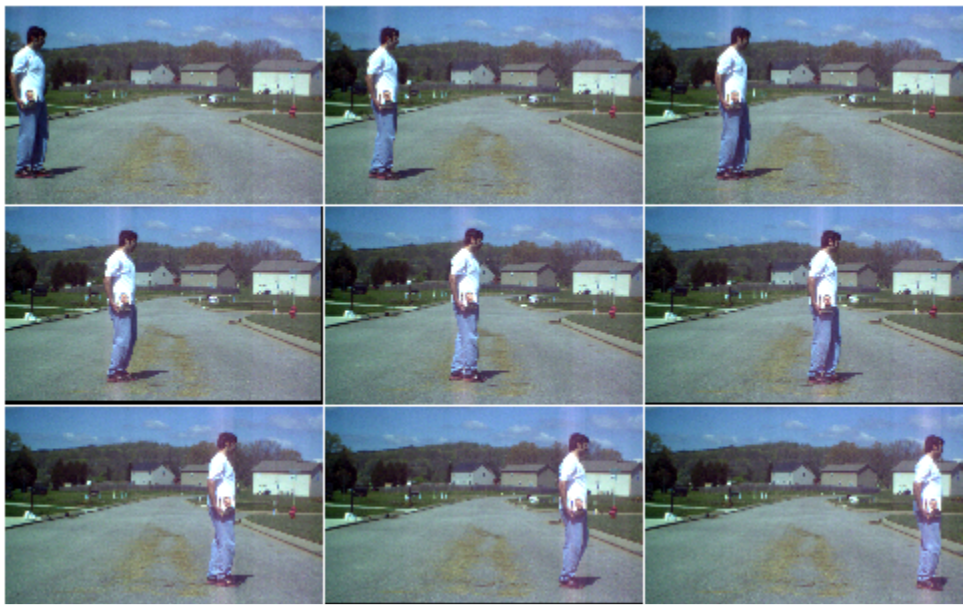


Figure 5.9: Test video example 9 frames out of stabilized 100 frame-video (road surveillance video 1.1) 25% of each dimension (original 400x640).

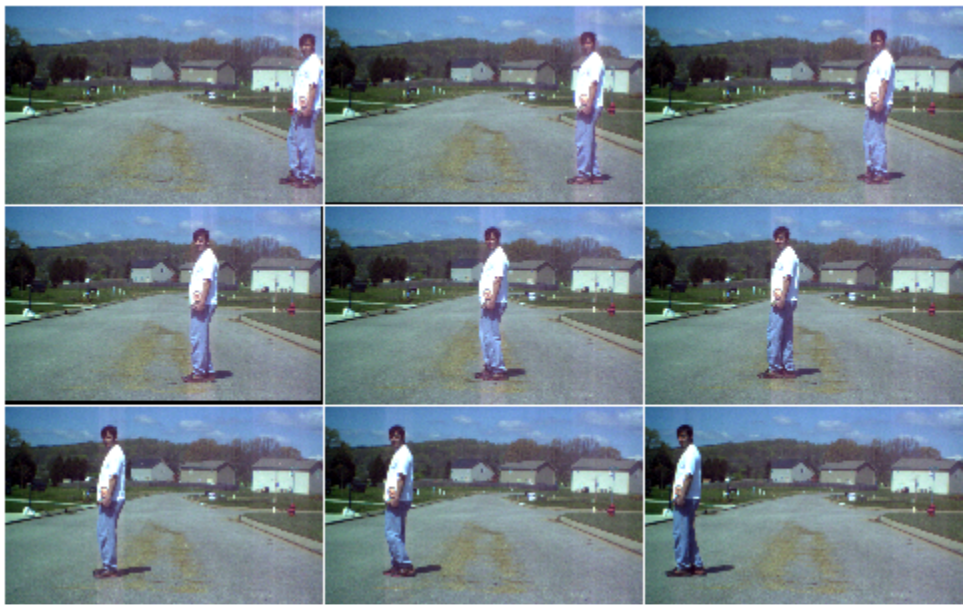


Figure 5.10: Test video example 9 frames out of stabilized 110 frame-video (road surveillance video 1.2) 25% of each dimension (original 400x640).

Figure 5.11: Test video example 9 frames out of stabilized 86 frame-video (road surveillance video 1.3) 25% of each dimension (original 400x640).

Figure 5.12: Test video example 9 frames out of stabilized 87 frame-video (road surveillance video 1.4) 25% of each dimension (original 400x640).

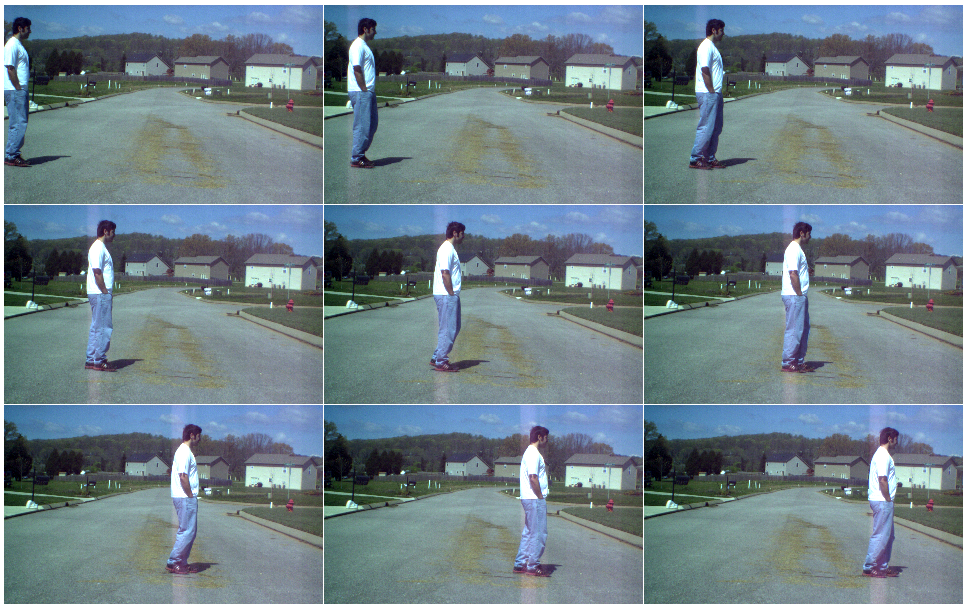


Figure 5.13: Test video example 9 frames out of 100 frame-video (road surveillance video 2.1) 25% of each dimension (original 400x640)

Figure 5.14: Test video example 9 frames out of 110 frame-video (road surveillance video 2.2) 25% of each dimension (original 400x640).

Figure 5.15: Test video example 9 frames out of 86 frame-video (road surveillance video 2.3) 25% of each dimension (original 400x640).

Figure 5.16: Test video example 9 frames out of 87 frame-video (road surveillance video 2.4) 25% of each dimension (original 400x640).

We deal with dynamic scenes with stationary camera system at the current state of our framework. Up to the scenario we put original images sequences ‘road surveillance video 1.1’, ‘road surveillance video 1.2’, ‘road surveillance video 1.3’, ‘road surveillance video 1.4’ are stabilized using a global motion suppression scheme. On the other hand images sequences ‘road surveillance video 2.1’, ‘road surveillance video 2.2’, ‘road surveillance video 2.3’, and ‘road surveillance video 2.4’ have only locally moving objects and they are not degraded by camera motion. All the frames (100, 110, 86, 87; 100, 110, 86, 87) of two distinct groups (image sequences from stationary camera case, and stabilized image sequences for non-stationary camera case) are introduced to motion segmentation process which consist utilization of Information Complexity guided GMMs based Background Modeling. The background representations are given previously in Chapter 4.

We use all possible information coming from the images in background modeling. However, it is not feasible to use all of the LR frames to compute HR representation. The vast majority of the SR algorithms use a short sequence of LR input frames to produce a single super-resolved high-resolution output frame (the MISO case). The objective of SR addressed in this dissertation is to obtain an estimate of one HR frame at each time from available observations. The same techniques may, however, be applied to resolution enhancement of videos by using a shifting window of processed LR frames utilizing sliding window approach. Window determines the subset of low-resolution frames to be processed. The window is moved forward in time to produce successive super-resolved frames in the output sequence. Various approaches may be taken to determine the subset of low-resolution frames used to compute the HR frames corresponding to the start and end of the observed low-resolution image sequence. We run a visual test to pick a number of LR images, four sequential frames of each shot towards SR Image Reconstruction, yet the adequacy of the information we obtained after doing so cannot be evaluated. The information we gather from LR frames is important, however controlling the frame differences is more important. Four frames from 8 datasets used in SR Image Reconstruction are shown in Figure 5.17 -24.

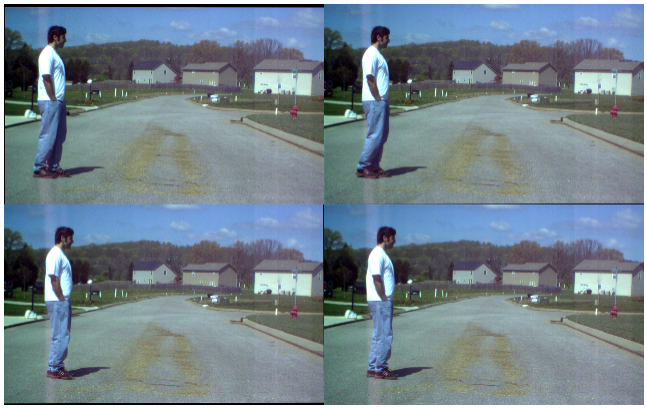


Figure 5.17: Frames #166, #167, #168, #169 of stabilized road surveillance video 1.1 to be used in SR Image Reconstruction, 25% of each dimension (original 400x640).

Figure 5.18: Frames #392, #393, #394, #395 of stabilized road surveillance video 1.2 to be used in SR Image Reconstruction, 25% of each dimension (original 400x640).

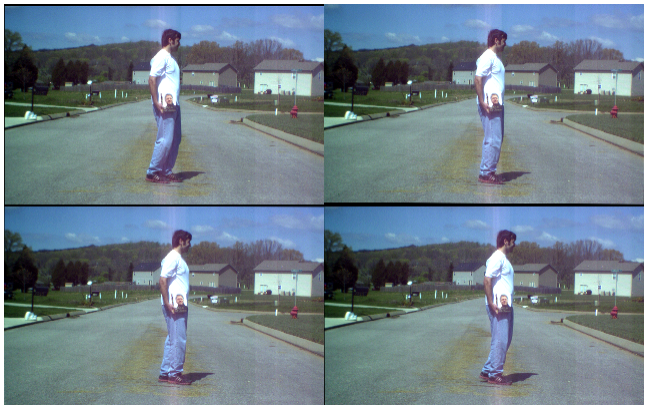


Figure 5.19: Frames #693, #694, #695, #696 of stabilized road surveillance video 1.3 to be used in SR Image Reconstruction, 25% of each dimension (original 400x640).

Figure 5.20: Frames #844, #845, #846, #847 of stabilized road surveillance video 1.4 to be used in SR Image Reconstruction, 25% of each dimension (original 400x640).

Figure 5.21: Frames #164, #165, #166, #167 of road surveillance video 2.1 to be used in SR Image Reconstruction, 25% of each dimension (original 400x640).

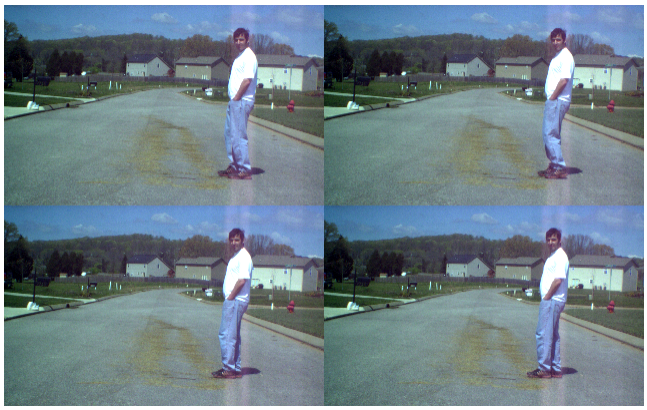


Figure 5.22: Frames #387, #388, #388, #390 of road surveillance video 2.2 to be used in SR Image Reconstruction, 25% of each dimension (original 400x640).

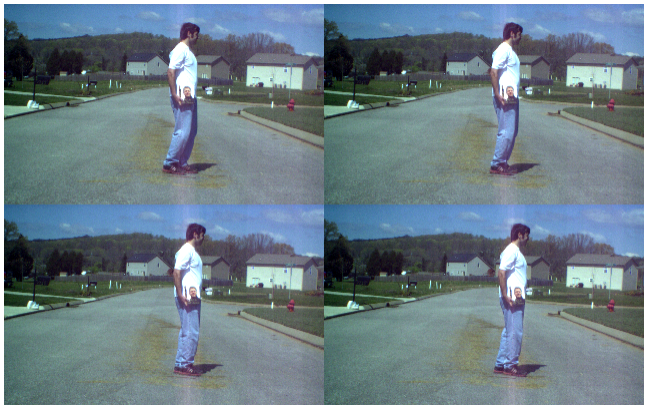


Figure 5.23: Frames #690, #691, #692, #693 of road surveillance video 2.3 to be used in SR Image Reconstruction, 25% of each dimension (original 400x640).

Figure 5.24: Frames #839, #840, #841, #842 of road surveillance video 2.4 to be used in SR Image Reconstruction, 25% of each dimension (original 400x640).

Using these 8 groups with four frames, and related background estimations, Foreground Detection known as background subtraction, simply thresholding the error between a model of the background without moving objects and the current image is applied. A post-processing to obtain the final silhouette of the foreground moving objects using multiple morphological operations and thresholding suppress false detections that are due to small motions in the background not captured by the model followed the background subtraction step. As the final step before SR Image Reconstruction rough local motion estimation of the segmented out moving objects is utilized to localize the information related to the moving regions. We extended the method in [Thevenaz98] with irregular shaped region of support. It is an automatic registration algorithm that minimizes the mean square intensity difference between a reference and a test dataset. It uses an explicit spline representation of the images in conjunction with spline processing, and is based on a coarse-to-fine iterative strategy (pyramid approach). The minimization they performed was a new variation of the Marquardt–Levenberg algorithm for nonlinear least-square optimization. In this study we restricted the geometric deformation model to rigid-body motion (rotation

and translation). The results for background subtraction and local motion compensation are given in Figure 5.25-40.

Figure 5.25: Segmented out regions of frames #166, #167, #167, #169 of stabilized road surveillance video 1.1, using GMMs based Background Modeling, (80% in each dimension).

Figure 5.26: Rough registration of the segmented out regions of frames #166, #167, #167, #169 of stabilized road surveillance video 1.1, (80% in each dimension).

Figure 5.27: Segmented out regions of frames #392, #393, #394, #395 of stabilized road surveillance video 1.2, using GMMs based Background Modeling, (80% in each dimension).

Figure 5.28: Rough registration of the segmented out regions of frames #392, #393, #394, #395 of stabilized road surveillance video 1.2, using [Thévenaz98], (80% in each dimension).

Figure 5.29: Segmented out regions of frames #693, #694, #695, #696 of stabilized road surveillance video 1.3, using GMMs based Background Modeling, (80% in each dimension).

Figure 5.30: Rough registration of the segmented out regions of frames #693, #694, #695, #696 of stabilized road surveillance video 1.3, (80% in each dimension).

Figure 5.31: Segmented out regions of frames #844, #845, #846, #847 of stabilized road surveillance video 1.4, using GMMs based Background Modeling, (80% in each dimension).

Figure 5.32: Rough registration of the segmented out regions of frames #844, #845, #846, #847 of stabilized road surveillance video 1.4, (80% in each dimension).

Figure 5.33: Segmented out region of frames #164, #165, #166, #167 of road surveillance video 2.1, using GMMs based Background Modeling, (80% in each dimension).

Figure 5.34: Rough registration of the segmented out regions of frames #164, #165, #166, #167 of road surveillance video 2.1, (80% in each dimension).

Figure 5.35: Segmented out region of frames #387, #388, #388, #390 of road surveillance video 2.2, using GMMs based Background Modeling, (80% in each dimension).

Figure 5.36: Rough registration of the segmented out regions of frames #387, #388, #388, #390 of road surveillance video 2.2, using, (80% in each dimension).

Figure 5.37: Segmented out region of frames #690, #691, #692, #693 of road surveillance video 2.3, using computed GMMs based Background Modeling, (80% in each dimension).

Figure 5.38: Rough registration of the segmented out regions of frames #690, #691, #692, #693 of road surveillance video 2.3, using, (80% in each dimension).

Figure 5.39: Segmented out region of frames #839, #840, #841, #842 of road surveillance video 2.4, using GMMs based Background Modeling, (80% in each dimension).

Figure 5.40: Rough registration of the segmented out regions of frames #839, #840, #841, #842 of road surveillance video 2.4, (80% in each dimension).

When the low-resolution images are roughly registered the samples of the different images can be combined to reconstruct a high resolution image. We employed following SR Image Reconstruction methods, and the results are given in Fig 5.41-48,

- Interpolation
- Papoulis-Gerchberg [Papoulis77]
- Iterated Back Projection [Keren88]

- Robust Super Resolution[Zomet01]
- POCS [Patti97]
- Structure-Adaptive Normalized Convolution [Pham06]
- Kriging [Krige51].

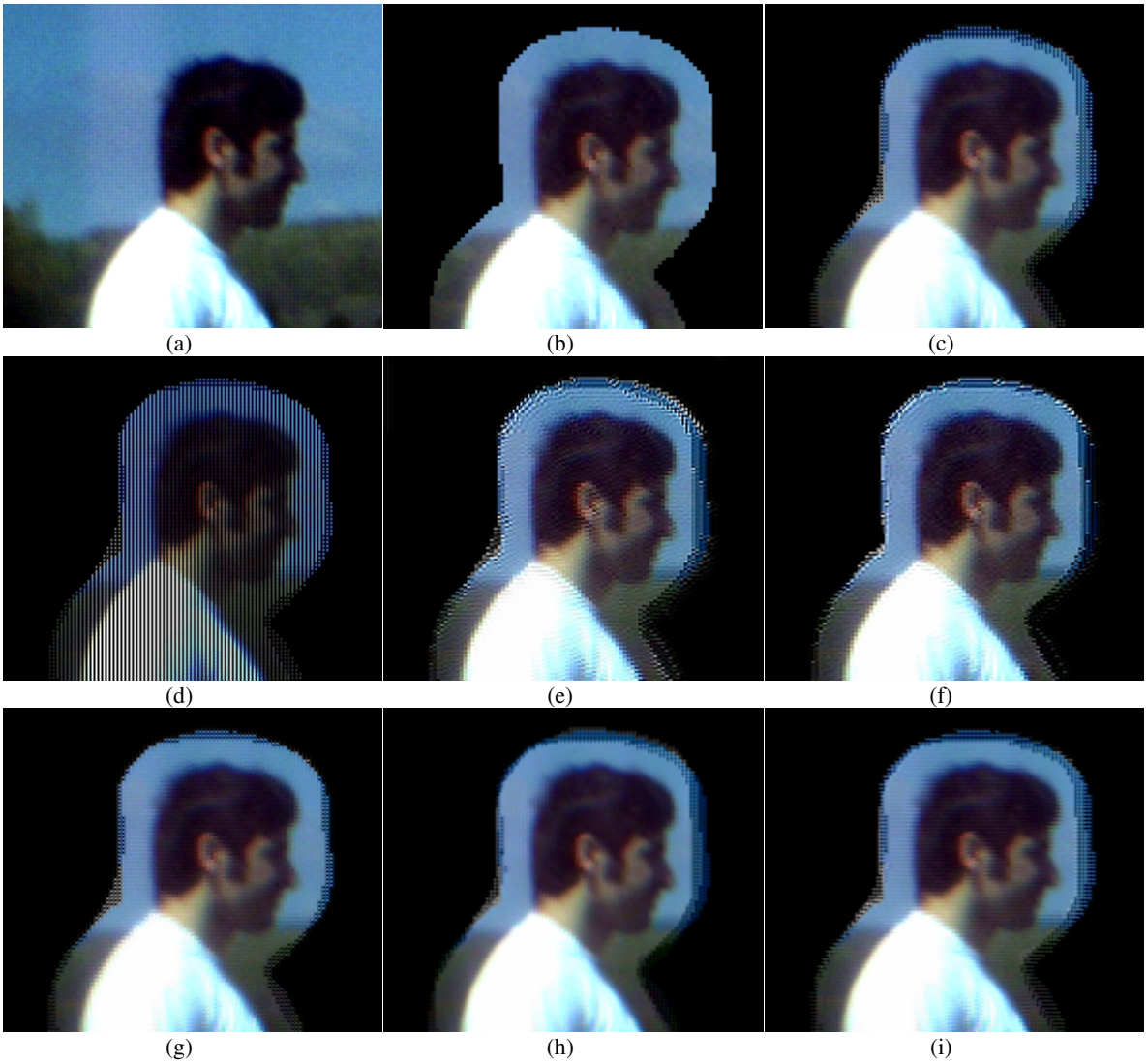


Figure 5.41: a) HR image, b) LR (interpolated) representations; SR Image Reconstruction using sub-pixel image registration [Vandawalle06] and several reconstruction methods, c) Interpolation, d)Papoulis-Gerchberg, e) Iterated back projection, f) Robust Super Resolution, g) Projection Onto Convex Sets, h) Structure Adapted Normalized Convolution, i) Kriging, for stabilized road surveillance video 1.1, (Images are cropped from the original size super-resolved HR output images).

Figure 5.42: a) HR image, b) LR (interpolated) representations; SR Image Reconstruction using sub-pixel image registration [Vandawalle06] and several reconstruction methods, c) Interpolation, d)Papoulis-Gerchberg, e) Iterated back projection, f) Robust Super Resolution, g) Projection Onto Convex Sets, h) Structure Adapted Normalized Convolution, i) Kriging, for stabilized road surveillance video 1.2, (Images are cropped from the original size super-resolved HR output images).

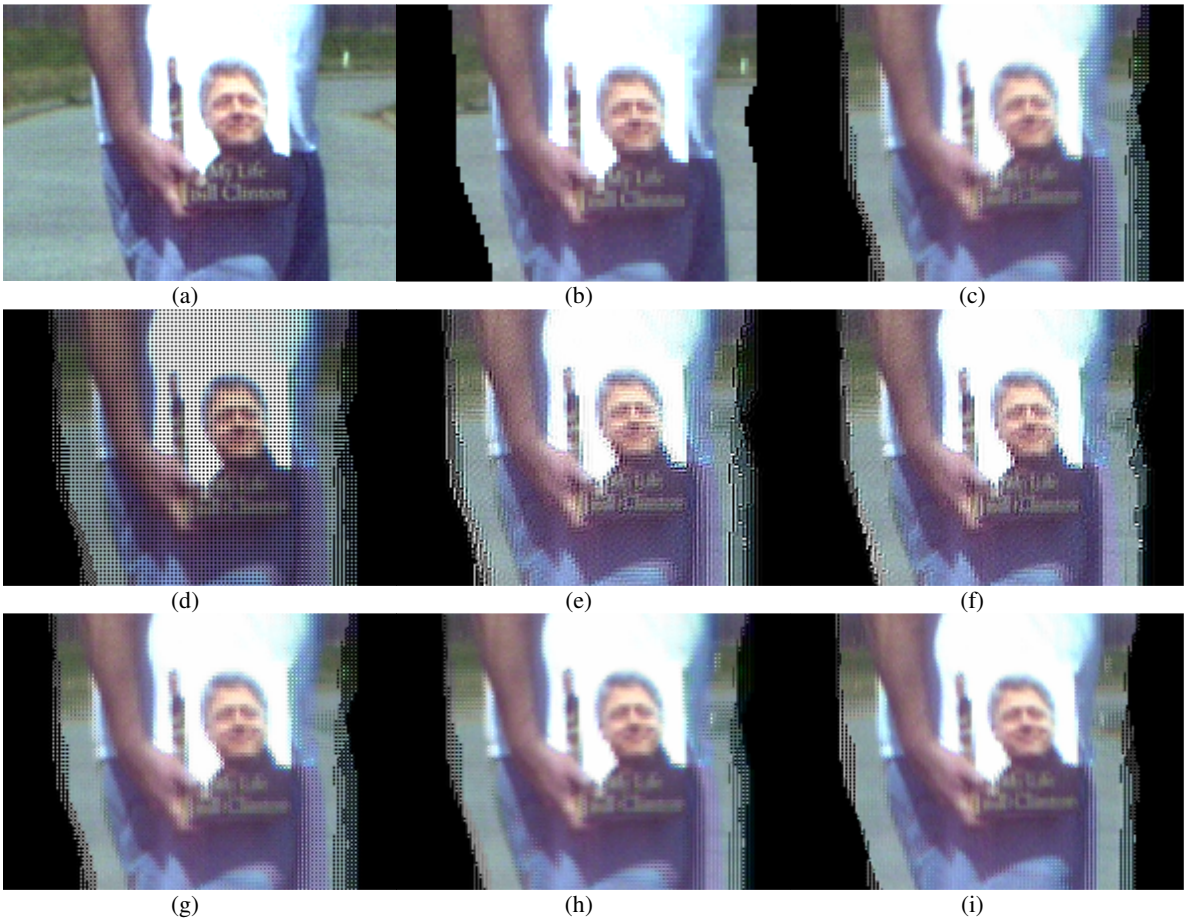


Figure 5.43: a) HR image, b) LR (interpolated) representations; SR Image Reconstruction using sub-pixel image registration [Vandawalle06] and several reconstruction methods, c) Interpolation, d)Papoulis-Gerchberg, e) Iterated back projection, f) Robust Super Resolution, g) Projection Onto Convex Sets, h) Structure Adapted Normalized Convolution, i) Kriging, for stabilized road surveillance video 1.3, (Images are cropped from the original size super-resolved HR output images).

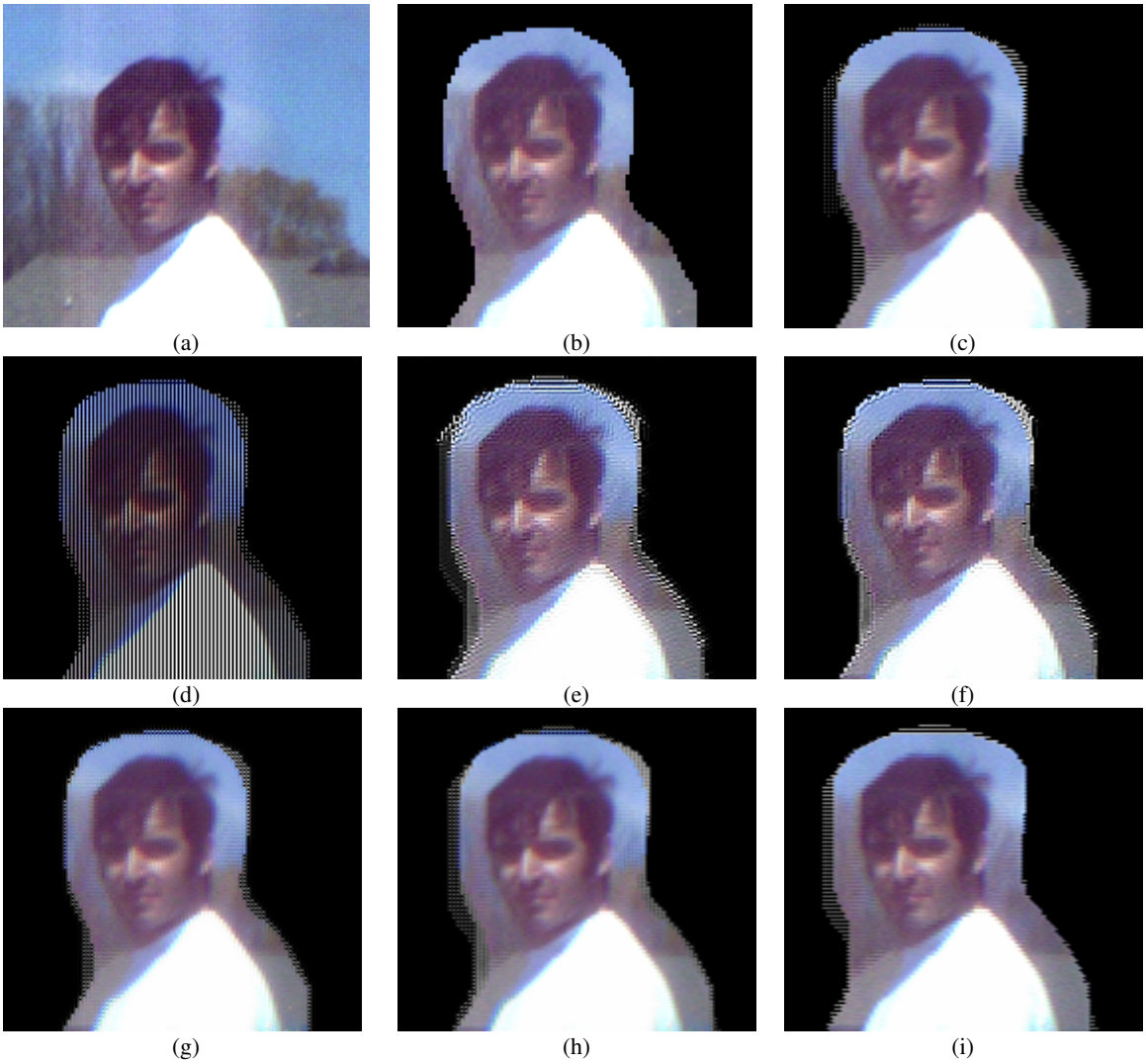


Figure 5.44: a) HR image, b) LR (interpolated) representations; SR Image Reconstruction using sub-pixel image registration [Vandawalle06] and several reconstruction methods, c) Interpolation, d)Papoulis-Gerchberg, e) Iterated back projection, f) Robust Super Resolution, g) Projection Onto Convex Sets, h) Structure Adapted Normalized Convolution, i) Kriging,, for stabilized road surveillance video 1.4, (Images are cropped from the original size super-resolved HR output images).

Figure 5.45: a) HR image, b) LR (interpolated) representations; SR Image Reconstruction using sub-pixel image registration [Vandawalle06] and several reconstruction methods, c) Interpolation, d)Papoulis-Gerchberg, e) Iterated back projection, f) Robust Super Resolution, g) Projection Onto Convex Sets, h) Structure Adapted Normalized Convolution, i) Kriging, for road surveillance video 2.1, (Images are cropped from the original size super-resolved HR output images).

Figure 5.46: a) HR image, b) LR (interpolated) representations; SR Image Reconstruction using sub-pixel image registration [Vandawalle06] and several reconstruction methods, c) Interpolation, d)Papoulis-Gerchberg, e) Iterated back projection, f) Robust Super Resolution, g) Projection Onto Convex Sets, h) Structure Adapted Normalized Convolution, i) Kriging, for road surveillance video 2.2, (Images are cropped from the original size super-resolved HR output images).

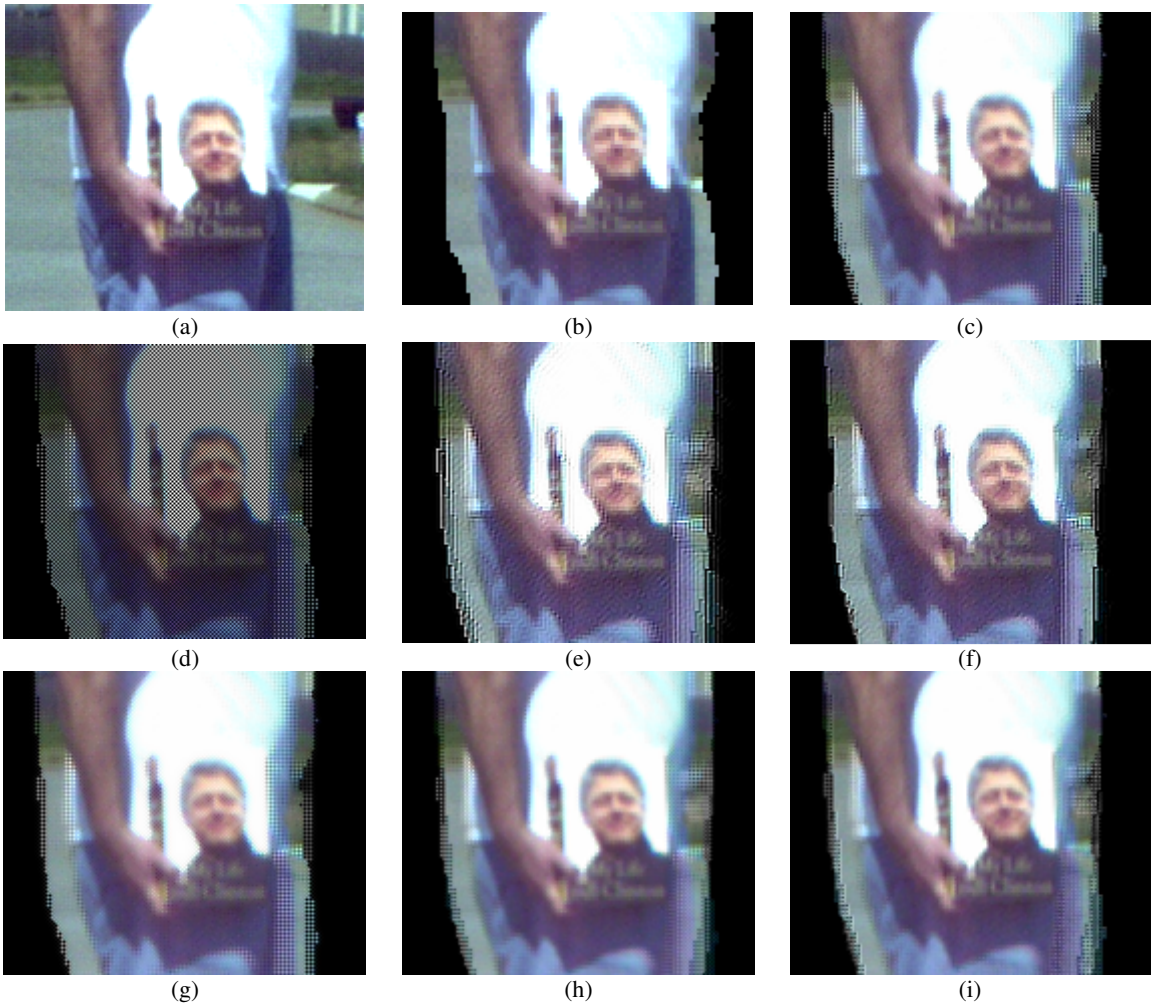


Figure 5.47: a) HR image, b) LR (interpolated) representations; SR Image Reconstruction using sub-pixel image registration [Vandawalle06] and several reconstruction methods, c) Interpolation, d)Papoulis-Gerchberg, e) Iterated back projection, f) Robust Super Resolution, g) Projection Onto Convex Sets, h) Structure Adapted Normalized Convolution, i) Kriging, for road surveillance video 2.3. (Images are cropped from the original size super-resolved HR output images).

Figure 5.48: a) HR image, b) LR (interpolated) representations; SR Image Reconstruction using sub-pixel image registration [Vandawalle06] and several reconstruction methods, c) Interpolation, d)Papoulis-Gerchberg, e) Iterated back projection, f) Robust Super Resolution, g) Projection Onto Convex Sets, h) Structure Adapted Normalized Convolution, i) Kriging, for road surveillance video 2.4. (Images are cropped from the original size super-resolved HR output images.

5.6 Summary

We presented a framework of Super Resolution Image/Video Reconstruction for the extracted regions (related to moving objects) we gathered from the previous block of motion segmentation process, in which we are having high level of displacements of the objects resulting from not only the local motion of the objects but the global motion of non-stationary imaging system. We utilized a frequency domain sub pixel image registration method to register a set of low-resolution, images representing just the moving regions in the scene. Planar rotation and translation parameters are precisely estimated by the method from [Vandewalle06]. After the sub-pixel image alignment, several interpolation techniques to closely accumulated information from the LR images are applied in order to reconstruct the HR representation of moving objects. It is proven that our framework with all the efforts previously to segment out and to localize the moving regions has a great impact on the last step SR Image Reconstruction algorithm.

6 Conclusions

In summary, we addressed Super Resolution (SR) Image Reconstruction framework with a focus on the task of Information Complexity guided Gaussian Mixture Models (GMMs) for Statistical Background Modeling which is used in motion segmentation. High level of local displacements of the moving objects imposed on the global instabilities arising from the non-stationary camera system is a bottleneck in the state of art SR methods. Contrary to traditional SR approaches we employed several steps to handle some crucial challenges to accumulate corresponding information from highly displaced moving objects. We have stressed the use of motion segmentation which provides us the ability of both using irregular-shaped region of support for local motion estimation of the moving objects and suppressing the information coming from background to comfort the reconstruction stage of the framework.

These questions were at the core of our efforts:

- Can we model the background of the scene optimally to extract out the moving objects?
- Can we accurately accumulate all of the information coming from the moving objects on which global motion of the camera systems to have super-resolved representations?

We believe our efforts in this dissertation offer a good answer to these questions.

6.1 Dissertation Key Points

The key points for the foundation of this research are the following:

Information Complexity guided Statistical Background Modeling

We introduced a new technique; Information Complexity guided Statistical Background Modeling. Thus, we successfully employ GMMs, which are optimal w.r.t information complexity criteria, for background modeling. Regions highly occupied by moving objects are extracted optimally using parameter maps for component number and the shape of the components for each pixel. Moving objects are segmented out through background subtraction which utilizes the computed background model. This technique produces superior results to competing background modeling strategies.

Image Reconstruction of moving regions in non-stationary imaging systems

A framework of SR Image/Video Reconstruction of the moving objects, of which we are having high level of displacements, is developed. For dynamic scenes our assumption is that the images are different from each other due to not only the local motion of the objects but also the global motion of the scene imposed by non-stationary imaging system. In this framework, contrary to traditional SR approaches, we employed several steps to compute the HR representations of the moving objects. These steps are; suppression of the global motion trajectories imposed on the image sequence, motion segmentation accompanied by background subtraction to extract moving objects, localization of moving objects through suppression of the local motion trajectories of the segmented out regions, and super-resolving accumulated information

coming from multiple LR frames to reconstruct HR representations of the moving objects. In either case of stationary or non-stationary camera systems we intend to generate super-resolved representation of the moving objects rather than that of whole scene at the SR Image Reconstruction process. This results in a reliable offline SR Image Reconstruction tool which deals with several types of dynamic scene changes, compensates the impacts of camera systems, and brings data redundancy through removing the background information. The framework proved to be superior to the state of algorithms which put no significant effort for the dynamic scene recordings with non-stationary camera systems.

6.2 New Questions and Future Research

Of course this research while claiming an important place among the state of the art methods of SR Image/Video Reconstruction and background modeling does not pretend to 'solve' the problem in any definitive way.

For the global motion compensation task, the motions are computed using consecutive frames, and a more complex stabilization in which the first and the last frames are forced to come close to each other globally, is not pursued. Our methodology show a similarity to the one used in motion based video compression standards. A more global scheme to strike a balance between the range of the motion trajectories and the level of global displacements should be investigated. The global instabilities are assumed to be drawing closed paths and not diverging from a zero mean, yet for mobile camera systems neither a simple stabilization nor the estimation of a static background is achievable.

As the second gap of the study, we can discuss the computational time for GMMs based clustering of the $(m \times n)$ histories each of size $(N \times 3)$, where m and n are the size of the image, N is the frame number. This comes usually at the price of a time loss. Considering 100 frames of size 400x640, if we run each GMM based clustering in 10^{-4} (s) (with a computer processor of Intel Core i5 working at 2.4GHz), it takes around 0.7(h) to compose a background and this makes the current state of the algorithm not suitable for online applications such as traffic monitoring. We favored getting optimal Gaussian components number and the covariance matrix structure and sacrificed running an online system.

After having the optimal background modeling the task of background subtraction or Foreground Detection to seek the final silhouette of the moving objects using multiple morphological operations, is obsolete. These operators cannot be applied automatically and thus this task demands a user intervention. Other than morphological operators an automatic, general way to determine the moving object boundaries should be pursued.

Bibliography

[Acharya05] T. Acharya, and A. K. Ray, *Image processing: principles and its applications*, John Wiley & Sons, (Hoboken, NJ), 2005.

[Akaike73] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle", *Second International Symposium on Information Theory*, pp.267–281, Budapest, 1973.

[Amintoosi07] M. Amintoosi, F. Farbiz, M. Fathy, M. Analoui, and N. Mozayani "QR decomposition-based algorithm for background subtraction", *ICASSP 2007*, vol.1, pp.1093-1096, April 2007.

[Baker02] S. Baker and T. Kanade, "Limits on super resolution and how to break them", *IEEE Transaction on*, *Pattern Analysis and Machine Intelligence,* vol. 204, no.9, September 2002.

[Baker03] S. Baker, R. Gross, T. Ishikawa and I. Matthews, "Lucas-Kanade 20 Years On: A Unifying Framework: Part 2", *International Journal of Computer Vision*, vol. 56, pp. 221-255, 2003.

[Banfield93] J. D. Banfield, and A. E. Raftery, "*Model-based Gaussian and non-Gaussian clustering*", Biometrics, vol.49, pp.803–821, 1993.

[Barnes11] R. Barnes, "Variogram Tutorial", *Technical Report : Golden Software , Inc.*, 2011.

[Black96] M. J. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," *International Journal of Computer Vision*, vol. 19, no. 1, pp. 57-91, 1996.

[Borman98] S. Borman, and R. L. Stevenson, "Super Resolution from Image Sequences - A review", *in Proc. of the 1998 Midwest Symposium on Circuits and Systems,* pp. 374-378*,* 1998.

[Borman99] S. Borman, and R. L. Stevenson, "Simultaneous multi-frame MAP Super Resolution video enhancement using spatio-temporal priors", *in Proc. of the IEEE International Conference on Image Processing (ICIP 99),* pp. 469-473, 1999.

[Borman02] S. Borman, and R. L. Stevenson, "Image Sequence Processing", *in R. G. Driggers editor, Dekker Encyclopedia of Optical Engineering*, pp.840-879, 2002.

[Bouwmans10] T. Bouwmans, F. El- Baf, and B. Vachon, "Statistical Background Modeling for Foreground Detection: A Survey", *Handbook of Pattern Recognition and Computer Vision, World Scientific Publishing*, vol.4, part 2, ch.3, pp.181-199, January 2010.

[Bovik09] A. C. Bovik, (ed.), *The essential guide to video processing*, Academic Press, 2009.

[Bozdogan88] H. Bozdogan, "ICOMP: A New Model-Selection Criteria", *In Bock, H. H.,* (ed.), *Classification and Related Methods of Data Analysis*, pp.599-608, Elsevier Science, (North-Holland), 1988.

[Bozdogan94] H. Bozdogan, "Mixture-Model Cluster Analysis Using Model Selection Criteria and a New Informational Measure of Complexity", *in Proc. of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, vol. 2, pp. 69–113, Dordrecht, the Netherlands. 1994.

[Bozdogan10] H. Bozdogan, "A new class of information complexity (ICOMP) criteria with an application to customer profiling and segmentation", *Istanbul University Journal of the School of Business Administration*, vol.39, pp.370–398, 2010.

[Celeux95] G. Celeux, and G. Govaert, "Gaussian parsimonious clustering models", *Pattern Recognition*, vol.28, no.5, pp.781-793, 1995.

[Chaudhuri01] S. Chaudhuri (Ed.), *Super Resolution Imaging*, Kluwer Academic Press, Boston, 2001.

[Chen07] Y. Chen, C. Chen, C. Huang, and Y. Hung, "Efficient hierarchical method for Background Subtraction", *Pattern Recognition*, vol.40, no.10, pp. 2706-2715, 2007.

[Cheng06] J. Cheng, J. Yang, Y. Zhou, and Y. Cui. "Flexible background mixture models for foreground segmentation", *Image and Vision Computing,* vol. 24, no.5, pp.473-482, 2006.

[Cheung05] S. Cheung, and C. Kamath, "Robust Background Subtraction with foreground validation for Urban Traffic Video", *Appl Signal Proc, Special Issue on Advances in Intelligent Vision Systems: Methods and Applications* (EURASIP 2005), New York, USA, vol.14, pp.2330-2340, 2005.

[Cucchiara03] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting Moving Objects, Ghosts, and Shadows in Video Streams", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.25, no.10, pp.1337-1342, 2003.

[El-Baf08] F. El-Baf, T. Bouwmans, and B. Vachon, "Fuzzy integral for moving object detection" in Proc. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2008)*, pp. 1729-1736 Hong-Kong, China, June 2008.

[Elgammal00] A. Elgammal, D. Harwood, and L.S. Davis, "Non-parametric model for Background Subtraction," *in Proc. of the Sixth* European Conference on Computer Vision (ECCV '00), vol.2, pp.751-767, 2000.

[Elgammal01] A. Elgammal, R. Duraiswami, and L. S. Davis ,"Efficient Non-Parametric Adaptive Color Modeling Using Fast Gauss Transform", *in Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'01),* vol. 2, pp.563-570, 2001.

[Elhabian08] S. Elhabian, K. El-Sayed, and S. Ahmed, "Moving object detection in spatial domain using background removal techniques - State-of-Art", *Recent Patents on Computer Science*, vol.1, no.1, pp. 32-54, 2008.

[Erar11] B. Erar, *Mixture model cluster analysis under different covariance structures using information complexity*, Master's Thesis, University of Tennessee, 2011.

[Fan06] C. Fan, J. Zhu, J. Gong, and C. Kuang, "POCS Super Resolution sequence Image Reconstruction based on improvement approach of Keren registration method", *in Proc. of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA '06),* pp. 333-337, 2006.

[Fitzgibbon03] A.W. Fitzgibbon, "Robust registration of 2D and 3D point sets", *Image and Vision Computing,* vol. 21, pp.1145-1153, 2003.

[Fraley98] C. Fraley, "Algorithms for model-based Gaussian hierarchical clustering", *SIAM Journal on Scientific Computing*, vol.20, pp.270-281, 1998.

[Fraley 02] C. Fraley, and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation", *Journal of the American Statistical Association*, vol.97, no.458, pp. 611–631, 2002.

[Francois99] A. R. Francois, and G. G. Medioni, "Adaptive color Background Modeling for real-time segmentation of video streams*", in Proc. of the Proceedings of the International Conference on Imaging Science, Systems, and Technology*, Las Vegas, NA, pp.227-232, June 1999.

[Friedman97] N. Friedman, and S. Russell, "Image segmentation in video sequences: A probabilistic approach," *In Proc. of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp.175-181, Aug. 1-3, 1997.

[Glasner09] D. Glasner, S. Bagon, and M. Irani, "Super Resolution from a Single Image", *in Proc. of the IEEE International Conference on Computer Vision (ICCV'09),* pp. 349 –356, 2009.

[Greiffenhagen01] M. Greiffenhagen, V. Ramesh, and H. Niemann, "The systematic design and analysis cycle of a vision system: A case study in video surveillance", *in Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'01),* 2001.

[Grinstead07] B. Grinstead, *Detail Enhancing Denoising of Digitized 3D Models from a Mobile Scanning System* , PhD Dissertation, University of Tennessee, 2007.

[Hardie97] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint MAP registration and high-resolution image estimation using a sequence of under-sampled images", *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1621–1633, 1997.

[Harville01] M. Harville, G. Gordon, and J. Woodfill, "Foreground Segmentation Using Adaptive Mixture Models in Color and Depth", *in Proc. of the IEEE Workshop on Detection and Recognition of Events in Video (EVENT'01),* pp.3-12, 2001.

[Hsu04] Hsu J. T., Yen C. C., Li C. C., Sun M., Tian B., and Kaygusuz M., "Application of wavelet-based POCS super resolution for cardiovascular MRI image enhancement", *in Proc. Third International Conference on Image and Graphics*, pp. 572 -575, Hong Kong, China, 2004.

[Irani91] M. Irani, S. Peleg, "Improving resolution by Image Registration", "*CVGIP: Graphical Models and Image Processing*", vol.53, no.3, pp.231-239, 1991.

[Irani94] M. Irani, B. Rousso, and S. Peleg, "Computing Occluding and Transparent Motions", *International Journal of Computer Vision*, vol.12, no.1, pp. 5-16, 1994.

[Irani98] M. Irani and P. Anandan, "Robust multi-sensor image alignment", *in Proc. of the IEEE International Conference on Computer Vision, (ICCV'98)*, Bombay, January, 1998.

[Irani00] M. Irani and P. Anandan, "About Direct Methods", *Vision Algorithms: Theory and Practice, (Book chapter), Lecture Notes in Computer Science*, vol.1883, pp.267-277, 2000.

[Jain 88] A. K. Jain, and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, (Englewood Cliffs, NJ), 1988.

[Javed02] O. Javed, K. Shafique, and M. Shah "A hierarchical approach to robust Background Subtraction using color and gradient information", *in Proc. of the IEEE Workshop on Motion and Video Computing (WMVC 2002)*, pp.22-27 2002.

[Julio05] C. S. Julio, C. R. Jung, and S. R. Musse, "Background Subtraction and Shadow Detection in Grayscale Video Sequences", *in Proc. of the 18th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI'05)*, pp. 189-196, 2005.

[Katsaggelos07] A. K. Katsaggelos, R. Molina and J. Mateos, *Super Resolution of Images and Video*, Morgan & Claypool Publishers, 2007.

[Kentaro99] T. Kentaro, K. John, B. Barry, and M. Brian, "Wallflower: Principles and Practice of Background Maintenance", *in Proc. of the IEEE International Conference on Computer Vision, (ICCV'99)* pp.255-261, 1999.

[Keren88] D. Keren, S. Peleg, and R. Brada, "Image sequence enhancement using sub-pixel displacements", *in Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'88),* pp.742-746, 1988.

[Khan06] E.A. Khan, A.O. Akyuz, and E. Reinhard, "Ghost Removal in High Dynamic Range Images", *in Proc. of the IEEE International Conference on Image Processing (ICIP'06),* pp.2005-2008, 2006.

[Kim05] K. Kim, T.H. Chalidabhongse, D. Harwood, and L.S. Davis, "Real-time foreground-background segmentation using codebook model", *Real-Time Imaging*, vol.11, pp.172-185, 2005.

[Kim07] H. Kim, R. Kitahara, I. Sakamoto, T. Toriyama, and K. Kogure, "Robust silhouette extraction technique using Background Subtraction", *10th Meeting on Image Recognition and Understand (MIRU'07)*, Hiroshima, Japan, July 2007.

[Kim10] K. I. Kim, and Y. Kwon, "Single-image Super Resolution using sparse regression and natural image prior" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.32, no.6, pp.1127-1133, 2010.

[Koendrik84] J. J. Koendrik, "The Structure of Images", *Biological Cybernetics*, vol.50, pp.363-370, 1984.

[Koller94] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell, "Towards robust automatic traffic scene analysis in real-time", *in Proc. of the IEEE Conference on Decision and Control*, vol.4, pp.3776-3781, 1994.

[Krige51] D. G. Krige, "Statistical approach to some basic mine valuation problems on the Witwatersrand", *Journal of the Chemical Metallurgical and Mining Society* , vol. 52, no. 6, pp. 119-139, 1951.

[Kuhne02] G. Kuhne, *Motion-based Segmentation and Classification of Video Objects*, PhD dissertation, University of Mannheim, 2002.

[Kullbak51] A. Kullback, and R. Leibler, "*On Information and Sufficiency*", *Annals of Mathematical Statistics*, vol.22, pp.79-86, 1951.

[Kumar00]P. Kumar, K. Sengupta and S. Ranganath, "Real time detection and recognition of human profiles using inexpensive desktop cameras", *in Proc. of the IEEE International Conference on Pattern Recognition (ICPR'00)*, pp.1096-1099, 2000.

[Lee03] D.S. Lee, J.J. Hull, and B. Erol, "A Bayesian framework for Gaussian mixture Background Modeling," *in Proc. of the IEEE International Conference on Image Processing (ICIP'03)*, vol.3, pp. 973-976, 2003.

[Li07] D. Li, R. M. Mersereau, and S. Simske, "Atmospheric Turbulence Degraded Image Restoration Using Principal Components Analysis", *IEEE Geoscience and Remote Sensing Letters*, vol.4, no.3, pp.340-344, 2007.

[Liyakathunisa09] Liyakathunisa, and V. K. Ananthashayana, “Super resolution blind reconstruction of low resolution images using wavelets based fusion”, *International Journal of Computer, Information and Systems Science and Engineering*, vol.2, no.2, pp. 106-110, 2009.

[Lucas81] B. D. Lucas, and T. Kanade, “An iterative image registration technique with an application to stereo vision”, *in Proc. of the Seventh International Joint Conference on Artificial Intelligence (IJCAI’81),* pp.674-679, 1981.

[Marcel97] B. Marcel, M. Briot, and R. Murrieta, “Calcul de translation et rotation par la transformation de Fourier”, *Traitement du Signal*, vol.14, no.2, pp.135-149, 1997.

[Mas03] J. Mas, G. Fernandez, "Video Shot Boundary Detection based on Color Histogram", in Proc. of the TRECVid Workshop, pp.28-38, 2003.

[McFarlane95] N. J. B. McFarlane, and C. P. Schofield, “Segmentation and tracking of piglets in images”, *Machine Vision and Applications*, vol.8, pp.187-193, 1995.

[Morellas03] V. Morellas, I. Pavlidis, and P. Tsiamyrtzis “DETER: Detection of events for threat evaluation and recognition”, *Machine Vision and Applications,* vol.15, pp.29-45, 2003.

[Nadimi04] S. Nadimi, and B. Bhanu “Physics-based cooperative sensor fusion for moving object detection”, in *Proc. of the IEEE Workshop on Computer Vision and Pattern Recognition (CVPRW’04)*, pp.108-115, 2004.

[Nguyen00] N. Nguyen and P. Milanfar, “A wavelet based interpolation restoration method for super resolution”, *Circuits Systems Signal Processing*, vol.19, pp.321-338, 2000.

[Okayama99] H. Okayama, and L. Wang, “Spatial Coherence Degradation of Light Influenced by Temperature and Aerosol by Use of Atmospheric Turbulence Chamber”, *Remote Sensor Environment,* vol. 69, pp.189-193*,* 1999.

[Papathanassiou05] C. Papathanassiou, and M. Petrou, “Super resolution: an overview”, *in Proc. of Int. Geoscience and Remote Sensing Symposium (IGARSS’05)*, vol.8, pp.5655-5658, July 2005.

[Papoulis77] A. Papoulis, “Generalized sampling expansion,” *IEEE Transactions on Circuits Systems*, vol. 24, no. 11, pp. 652–654, 1977.

[Park03] 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.

[Patti97] A.J. Patti, M.I. Sezan, and A.M. Tekalp, “Super Resolution video reconstruction with arbitrary sampling lattices and nonzero aperture time”, *IEEE Trans. on Image Processing*, vol.6, no.8, pp.1064-1076, 1997.

[Pavlidis01] I. Pavlidis, V. Morellas, P. Tsiamyrtzis, and S. Harp, “Urban Surveillance Systems: From the Laboratory to the Commercial World”, *in the Proc. of the IEEE*, vol.89, no.10, pp.1478-1497, 2001.

[Pearson94] K. Pearson, “Contributions to the Mathematical Theory of Evolution”, *Philosophical Transactions of the Royal Society of London A*., Vol.185, pp. 71-110, 1894.

[Pham06] T. Q. Pham, L. J. van Vliet and K. Schutte, "Robust Fusion of Irregularly Sampled Data Using Adaptive Normalized Convolution", *EURASIP Journal on Applied Signal Processing*, 2006.

[Piccardi04] M. Piccardi, "Background Subtraction techniques: A review", in *Proc. of the International Conference on Systems, Man and Cybernetics (SMC'04),* pp.3199-3204, 2004.

[Reddy96] B. S. Reddy, B. and N. Chatterji, "An FFT-based technique for translation, rotation, and scale-invariant image registration," *IEEE Transactions on Image Processing*, vol.5, no.8, pp.1266-1271, 1996.

[Rhee99] S.H. Rhee and M.G. Kang, "Discrete cosine transform based regularized high-resolution Image Reconstruction algorithm," *Opt. Eng.,* vol.38, no.8, pp.1348-1356, 1999.

[Ridder95] C. Ridder, O. Munkelt, and H. Kirchner, "Adaptive Background Estimation and Foreground Detection using Kalman-Filtering", *in the Proc. of the International Conference on recent Advances in Mechatronics ,(ICRAM'95)*, pp.193-199,1995.

[Saddot95] D. Saddot, N.S. Kopeika, and S.R. Rotman, "Incorporation of Atmospheric Blurring Effects in Target Acquisition Modeling of Thermal Images", *Infrared Physics Technology,* vol. 36, no.2, pp.551-564, 1995.

[Schindler06] K. Schindler, and H. Wang, "Smooth foreground-background segmentation for video processing" in the Proc. of the *Asian Conference on Computer Vision (ACCV'06)* , pp.581-590, 2006.

[Schultz94] 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.333-342, 1994.

[Schultz96] 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, 1996.

[Schwarz78] G. Schwarz, "Estimating the Dimension of a Model", Analysis and Statistics", vol.6, pp.461-464, 1978.

[Shen07] H. Shen, L. Zhang, B. Huang, and P. Li, "A MAP approach for Joint Motion Estimation, Segmentation, and Super Resolution", *IEEE Transactions on Image Processing*, vol.16, no.2, pp.479-490, 2007.

[Sroubek07] F. Sroubek, G. Cristobal, and J. Flusser, "A unifed approach to super resolution and multi-channel blind deconvolution", *IEEE Transactions on Image Processing*, vol.16, no.9, pp.2322-2332, 2007.

[Stauffer99] C. Stauffer, and W.E.L. Grimson, "Adaptive background mixture models for real-time tracking", *in Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'99)*, vol.2, pp.246-252, 1999.

[Stenger01] B. Stenger, V. Ramesh, N. Paragios, F. Coetzee, and J. Buhmann, "Topology free hidden Markov models: Application to Background Modeling", *in Proc. of the IEEE International Conference on Computer Vision (ICCV'01)*, pp.294-301, 2001.

[Sun06] Y. Sun, B. Li, B. Yuan, Z. Miao, and C. Wan, "Better foreground segmentation for static cameras via new energy form and dynamic graph-cut", *in Proc. of the IEEE International Conference on Pattern Recognition (ICPR'06)*, pp.49-52, 2006

[Szeliski06] R. Szeliski, "Image Alignment and Stitching: A Tutorial", *Foundations and Trends in Computer Graphics and Computer Vision*, vol.2, no.1, pp.1-104, 2006.

[Tekalp95] A. M. Tekalp, *Digital Video Processing*, Prentice Hall, 1995.

[Thévenaz98] P. Thévenaz, U.E. Ruttimann, and M. Unser, "A Pyramid Approach to Subpixel Registration Based on Intensity*", IEEE Transactions on Image Processing*, vol. 7, no. 1, pp. 27-41, 1998.

[Thirion96] J. P. Thirion, "Non-rigid matching using demons", *in Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'96)*, pp.245-251, 1996.

[Tsai84] R. Y. Tsai, T. S. Huang, "Multi-frame Image Restoration and Registration", *Advances in Computer Vision and Image Processing*, vol.1, pp.317-339, 1984.

[Tuzel05] O. Tuzel, F. Porikli and P. Meer, "A Bayesian Approach to Background Modeling," in *Proc. of the IEEE Workshop on Computer Vision and Pattern Recognition (CVPRW'05)*, vol.3, pp.58-63, 2005.

[Vandawalle06] P. Vandawalle, S. Süsstrunk, and M. Vetterli, "A frequency domain approach to registration of aliased images with application to Super Resolution", *EURASIP Journal on Applied Signal Processing*, pp.1-14, 2006.

[Wang05] H. Wang, L. Dong, J. O'Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung, "Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy", *Phys. Med. Biol.*, vol.50, pp.2887-2905 , 2005.

[Walters07] B. D. Walters, and W. A. Clarke, "Comparison of two terrestrial atmospheric turbulence suppression algorithms", *in Proc. of Africon 2007*, pp.1-7, 2007.

[Wang04] Z. Wang and F. Qi, "On ambiguities in Super Resolution modeling", *IEEE Signal Processing Letters*, vol.11, no.8, pp.678-681, 2004.

[Wang05] D. Wang, W. Xie, J. Pei, and Z. Lu, "Moving area detection based on estimation of static background", *J. Inform. Comput. Science,* vol.2, no.1, pp.129-134, 2005.

[Wheeler07] F.W. Wheeler, X. M. Liu, and P. H. Tu, "Multi-frame Super Resolution for face recognition", *in the Proc. of the First IEEE International Conference on Biometrics: Theory, Applications, and Systems,* pp.1-6,. 2007.

[Willet03] R. Willet, I. Jermyn, R. Nowak, and J. Zerubia, "Wavelet based super resolution in astronomy", *In Proc. of Astronomical Data Analysis Software and Systems,* vol. 314, pp.107-116, 2003.

[Wren97] C. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: Real-Time Tracking of the Human Body", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol.19, no.7, pp.780-785, 1997.

[Yang09] J. Yang and D. Schonfeld, "New results on performance analysis of Super Resolution Image Reconstruction, *in Proc. of the IEEE International Conference on Image Processing (ICIP'09)*, pp.1517-1520, 2009.

[Yu08] J. Yu and B. Bhanu, "Super Resolution of deformed facial images in video", *in Proc. of the IEEE International Conference on Image Processing (ICIP'08)*, pp.1160-1163, 2008.

[Zhang06[Zhang H, and Xu D. “Fusing Color and Texture Features for Background Model”, *in the Proc. of the Third Int Conf on Fuzzy Systems and Knowledge Discovery*, pp. 887-893, 2006.

[Zibetti05] M.V.W. Zibetti, and J. Mayer, "Simultaneous Super Resolution for video sequences", *in Proc. of the IEEE International Conference on Image Processing (ICIP’05)*, vol.1, pp.11-14, Sept. 2005.

[Zivkovic04] Z. Zivkovic, “Improved adaptive Gaussian mixture model for Background Subtraction”, *in Proc. of the IEEE International Conference on Pattern Recognition (ICPR’04)*, pp.28-31, 2004.

[Zomet01] A. Zomet, A. Rav-Acha, and S. Peleg, “Robust super resolution,” *in Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR’01),* vol. 1, pp. 645-650, 2001.

Vita

Muharrem Mercimek was born in Kahramanmaras, Turkey. He attended Yildiz Technical University in Istanbul where he received a Bachelor of Science degree in 2001 and a Master of Science degree in 2003. He joined the Imaging, Robotics, and Intelligent Systems Laboratory (IRIS) at the University of Tennessee as a graduate research assistant. He earned his Doctor of Philosophy degree with a major in Electrical Engineering in May 2013.