Ph.D. in Electronic Engineering and Computer Science
Dept.of Electrical and Electronic Engineering
University of Cagliari, Sardinia, Italy

Picture Processing for Enhancement and Recognition

Ing. Tiziana Dessì

Advisor: Prof. Daniele Giusto

Curriculum: ING-INF/03 – Ph.D. in Electronic Engineering and Computer Science

XXV Cycle
a.a. 2011-2012

To Martina and Riccardo

Introduction

Recent years have been characterized by an incredible growth in computing power and storage capabilities, communication speed and bandwidth availability, either for desktop platform or mobile device. The combination of these factors have led to a new era of multimedia applications: browsing of huge image archives, consultation of online video databases, location based services and many other. Multimedia is almost everywhere and requires high quality data, easy retrieval of multimedia contents, increase in network access capacity and bandwidth per user.

To meet all the mentioned requirements many efforts have to be made in various research areas, ranging from signal processing, image and video analysis, communication protocols, etc.

The research activity developed during these three years concerns the field of multimedia signal processing, with particular attention to image and video analysis and processing. Two main topics have been faced: the first is relating to image and video reconstruction/restoration (using super resolution techniques) in web based application for multimedia contents' fruition; the second is relating to image analysis for location based systems in indoor scenario.

The first topic is relating to image and video processing, in particular the focus has been put on the development of algorithm for super resolution reconstruction of image and video sequences in order to make easier the fruition of multimedia data over the web. On one hand, latest years have been characterized by an incredible proliferation and surprising success of user generated multimedia contents, and also distributed and collaborative multimedia database over the web. This brought to serious issues related to their management and maintenance: bandwidth limitation and service costs are important factors when dealing with mobile multimedia contents’ fruition. On the other hand, the current multimedia consumer market has been characterized by the advent of cheap but rather high-quality high definition displays. However, this trend is only partially supported by the deployment of high-resolution multimedia services, thus the resulting disparity between content and display formats have to be addressed and older productions need to be either re-mastered or post-processed in order to be broadcasted for HD exploitation. In the presented scenario, super-resolution reconstruction represents a major solution. Image or video super resolution techniques allow restoring the original spatial resolution from low-resolution compressed data. In this way, both content and service providers, not to tell the final users, are relieved from the burden of providing and supporting large multimedia data transfer. In this context three works have been proposed. The first one is focused on super resolution techniques applied to huge database image browsing over the web. The proposed solution allows significant improvements of the service interactivity by increasing the image spatial resolution so that only thumbnail version of the images can be sent over the network. In the proposed work, the low-resolution image is first analyzed to identify several features that are significant for visual rendering and scene understanding. Such a classification is based on local frequency composition: uniform regions, edges and textures. The identified regions are then treated differently depending on the relative visual significance. Each region is further analyzed and a different interpolation approach is adopted, ranging from plain linear interpolation for homogeneous areas to edge area analysis and selective anisotropic interpolation. The combination of image region classification and adaptive-anisotropic interpolation is the main innovation of the proposed approach. The other two works deals with super-resolution applied to video sequences. The first solution is based on back projection and motion estimation. In particular, the high-resolution video sequence is reconstructed through iterative processing of inter-frame information and interpolative techniques. Resolution enhancement is based on iterative update of the high-resolution image estimate through motion and scene change detection. The second solution resorts to the use of the bilateral filtering. The proposed algorithm extends the use of the bilateral filter, traditionally used for still pictures, through the exploitation of the space-time domain and the development of edge-based samples estimation.

The second topic addressed during my Phd research activity is related to the implementation of an image based positioning system for an indoor navigator. As modern mobile device become faster, classical signal processing is suggested to be used for new applications, such location based service. The exponential growth of wearable devices, such as smartphone and PDA in general, equipped with embedded motion (accelerometers) and rotation (gyroscopes) sensors, Internet connection and high-resolution cameras makes it ideal for INS (Inertial Navigation System) applications aiming to support the localization/navigation of objects and/or users in an indoor environment where common localization systems, such as GPS (Global Positioning System), fail. Thus the need to use alternative positioning techniques.

In this context some image based positioning techniques have been investigated: the first one is based on plane homography and affine transformation while the second one is based on SURF. In the first case the considered scenario includes the presence of geo-referenced 2D-tags placed in some known, key positions of the site to be visited. By taking a photo of the tags, the system is able to initialize and subsequently re-calibrate the location data. To improve the calibration accuracy, the focus has been put on computing the exact position of the user (based on the known position of the tag) in terms of orientation and distance from the reference point using plane homography and affine transformation. This allows to correct perspective and projective distortion from the taken photo and derive information about the viewing angle (the user's orientation) and distance between camera and object. In the second work the smartphone's videocamera is used to identify known keypoints, named anchors previously identified and geo-referenced in the building map. For a periodic position fix, an image-based localization system is employed. By developing local feature detection, description and matching between a query image, acquired by the user with the built-in camera of the smartphone, and a database containing a collection of geo-referenced images related to the chosen environment, the user's position can be accurately fixed. The proposed solution is based on the SURF (Speed-up robust features), which allows for a quick and effective detection of image features without being affected by the user's viewpoint.

Published Papers

- Atzori Luigi, **Tiziana Dessi'**, Vlad Popescu “Indoor navigation system using image and sensor data processing on a smartphone” 13th International Conference on Optimization of Electrical and Electronic Equipment (OPTIM), Brasov, Romani, 24-26 May 2012

- Alberto Serra, **Tiziana Dessì**, Davide Carboni, Vlad Popescu, Luigi Atzori “Inertial Navigation Systems for User - Centric Indoor Applications” 3rd International NEM Summit Barcelona, Spain, October 13-15, 2010.

- Giaime Ginesu, **Tiziana Dessì**, Luigi Atzori, Daniele D. Giusto “Adaptive Bilateral Filtering For Superrsolution Reconstruction Of Video Sequences” 6th International ICST Mobile Multimedia Communications Conference (MOBIMEDIA 2010), September 2010.

- Giaime Ginesu, **Tiziana Dessì**, Luigi Atzori, Daniele D. Giusto “Super-Resolution reconstruction of video sequences based on back-projection and motion estimation," 5th International Mobile Multimedia Communications Conference (MobiMedia 2009), London, UK, September, 7-9, 2009

- Giaime Ginesu, **Tiziana Dessì**, Luigi Atzori, Daniele D. Giusto "Composite Interpolation Approach to Super-Resolution for Image Database Browsing over the Web," International Workshop on Content-Based Multimedia Indexing (CBMI08), pp. 315-322, London, UK, June, 18-20, 2008.

Chapter 1 Image and Video Processing Techniques

This section provides an overview of current technologies related to Image and Video Processing. Image processing is any form of signal processing for which the input is an image, such as photos or frames extracted from a video sequence; the output of image processing can be either an image or a set of characteristics or parameters related to the image. Usually the image is treated as a two-dimensional signal and standard signal-processing techniques is applied to it. Video processing is a particular case of signal processing: the input and the output signals are video sequence. Image and Video processing techniques are used in many science application: including medical imaging, satellite imaging, and video applications. Synthetic zooming of region of interest (ROI) is an important application in surveillance, forensic, scientific, medical, and satellite imaging. Another application is conversion from an NTSC video signal to an HDTV signal since there is a clear and present need to display a SDTV signal on the HDTV without visual artifacts.

In the following a description of the most common and used image and video processing techniques.

1.1 Reconstruction Techniques

The continuous development of image processing applications has increased the demand of high resolution images since they are not only more pleasant to look at, but they do provide a number of additional details that are important for the analysis of the images in a variety of applications. In fact, in most applications for image processing the high resolution (HR) is not only desired, but often necessary.

The term "Super Resolution" refers to the process by which it is possible to obtain a high-resolution image (HR) starting from one or more lower-resolution images (LR).

Super-resolution, also spelled as super resolution and superresolution, is a term for a set of methods of upscaling images or video.

The resolution of an image provides a measure of the quality of the image itself. The higher the resolution the greater the density of pixels (elementary points) that form the image, and then the more detailed that it contains. In the last twenty years there have been proposed many methods for super-resolution digital images. They are classified according to the number of LR images on which they work, thus we deals with Single-Frame and Multi-Frame super-resolution techniques. Most super-resolution techniques are based on the same idea: using information from several different images to create one upsized image. Algorithms try to extract details from every image in a sequence to reconstruct other frames. This multiframe approach differs significantly from sophisticated image (Single Input Single Output) upsizing methods which try to synthesize artificial details. The first infact, integrate information coming from frame slightly shifted in the same scene, while the latter (SISO) are essentially based on the process of interpolation of the pixels of the original image. A further distinction between Multi Frame techniques is related to the fact that we work on static images (Multiple Input Single Output) or video clips (Multiple Input Multiple Output). In both cases the improvement of the resolution takes place thanks to the fusion of the information contained in different frames of lower resolution and requires an accurate estimate of the relative motion between the latter and the frame taken as reference. About single frame techniques (Single Input Single Output), only few approaches have been proposed over the last few years. In these cases the improvement of the resolution takes place with or without the help of one or a set of training images extrapolated from scenes of the same type or of different type.

First work on this topic was published in 1984 [1] and the term "Super-resolution" itself appeared at around 1990 [2].

Methods usually discussed in scientific literature try to reproduce process of losing quality when shooting image/video with low-resolution cameras and then solve inverse problem of finding image/video which being downsized with that process gives us known low-resolution material. This is an ill-posed inverse problem which doesn't have straightforward solution and usually requires some additional regularization (applying some artificial constraints) and huge CPU time to check an awful lot of variants. Modern signal processing techniques, such as super resolution, are usually simpler but still effective. The major advantage of the signal processing approach is that it may cost less and the existing LR imaging systems can be still utilized.

Super-resolution (SR) works effectively when several low resolution images contain slightly different perspectives of the same object. Then total information about the object exceeds information from any single frame. The best case is when an object moves in the video. Motion detection and tracking are then employed to benefit upscaling. If an object doesn't move at all and is identical in all frames, no extra information can be collected. If it moves or transforms too fast then it looks very different in different frames and it's too hard to use information from one frame in reconstructing the other.

1.1.1 Super-resolution approaches

This section illustrates the main approaches that have been taken on to address the issue of image super-resolution. In the following, LR and HR refer to low-resolution and high-resolution images, respectively. The former represents the starting point of the signal processing procedure, whereas the latter is its output. The input LR image can be original image itself or a sub-sampled representation of the original image.

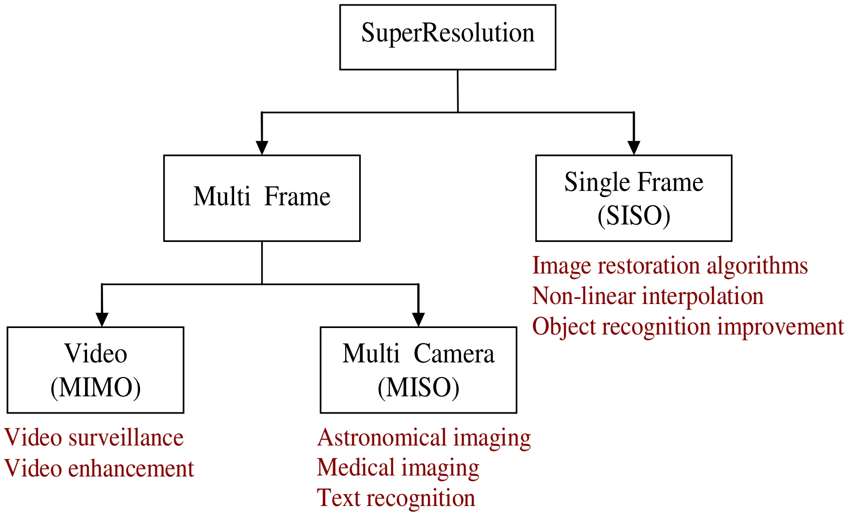


Fig. 1.1 Approaches to SuperResolution

The main classification of super-resolution approaches is based on the number of input images, so that we have Multi-Frame and Single-Frame techniques. In the first case, information coming from more frames is exploited whereas in the second case the output HR image is obtained from a single LR image. Multiframe techniques differ on the basis of whether the input data are static images (MISO - Multiple Input Single Output) or motion pictures (MIMO - Multiple Input Multiple Output). In both cases, the resolution improvement is obtained through the fusion of information coming from different frames at low resolution and requires accurate motion estimation operations.

1.1.1.1 Single Input Single Output Techniques

As to the techniques based on single-frame information (Single Input Single Output), image magnification is often obtained from one or more sets of training images coming from either similar or different scene kinds. The most common interpolation methods, such as bi-cubic and spline, estimate a continuous function that approximates the image signal on the basis of a set of local functions; the resulting function can then be resampled at the desired resolution. Some works propose the adoption of image sharpening operators to improve the image clarity at the expenses of additional artifacts [3]. To avoid such side effects, as jagged contours, [4] proposes a method attempting to produce smooth reconstructions of the image level curves while still preserving image fidelity. This is similar to other iterative reconstruction algorithms and to Bayesian restoration techniques, but instead of assuming smoothness prior for the underlying intensity function, it assumes smoothness of the level curves.

In [5], Schultz and Stevenson at first show that the image super-resolution problem is ill-posed and give a new definition. The problem is then solved by means of a Maximum A Posteriori (MAP) technique. Two techniques are presented: a constrained optimization approach for noise-free images and images affected by Gaussian noise. The image is described by means of a statistical model which includes the Huber convex function. In [6], the authors propose the use of the Markov Random Fields (MRF) to obtain the Bayes estimates; as a result, they obtain a model of realistic borders which is, at the same time, a viable MAP stable solution.

The method proposed in [7] suggests adopting different strategies for interpolation according to the characteristics of the region to be processed. The image is divided in blocks, which are then grouped to have uniform regions. On the basis of this classification, the algorithm applies either a multidirectional interpolation or a simple weighted average. The basic idea is to perform a weighted interpolation on the basis of the image gradient. The weight assigned to each pixel is set according to the following objectives: in the regions with sharp edges, the interpolation is performed parallel to the border direction; in flat regions the adjacent weights are quite similar so that the resulting effect is a simple moving average; in regions with mild edges the interpolator provides results in the middle with respect to the previous cases. A similar approach is proposed in [8]: the information about the discontinuity and the luminance variance drive the application of multidirectional interpolative techniques. More recently, a new approach bringing to learning-based methods has been proposed. It makes use of training sets analysis to acquire details of LR images. Afterwards, the learnt relations are used to predict realistic details in other images [9, 10]. Graphic models have also been proposed to obtain super-resolution images [11]. The flexibility of the proposed models incorporates the features of the natural images to define the compatibility of the HR image pixels that have to be estimated. Battiato *et al* [12] propose a method that takes into account information about discontinuities or sharp luminance variations while doubling the input picture. This is realized by a nonlinear iterative procedure of the zoomed image and could hence be implemented with limited computational resources. The algorithm works on monochromatic images and RGB color pictures.

In the last years, several works have been developed which use the gradient information to apply the best mask to each pixel being processed. For example, Rodrigues *et al.* [13] have proposed an adaptive edge-preserving algorithm which improves the level of the details and edges in the zoomed image through local threshold computation. Fu *et al.* [14] present a unified bidirectional flow process, where an inverse diffusion is performed to enhance edges along their normal directions, while a normal diffusion is performed with respect to the other directions. Martin *et al.* [15] apply the full color image reconstruction based on color filter array (CFA) zooming, CFA interpolation. Finally, Yan *et al.* [16] use a classical zooming algorithm based on gradient analysis of the input image only exploiting horizontal and vertical direction. All these approaches do not reduce the blurring effect and often present high computational complexity. In [17], Battiato *et al.* propose a gradient analysis algorithm that extends in some sense the original work they proposed in [12]. More precisely, they take into account the information about discontinuities or sharp luminance variations while increasing the input picture.

A novel algorithm that integrates bilateral filtering and back-projection is presented in [17]. The former achieves edge-preserving image smoothing while the latter minimizes the reconstruction error with an edge-based iterative procedure. In [18], the authors find the connection between the soft edge smoothness and a soft cut metric through a generalization of the Geocuts method. This term is incorporated into an objective function to produce smooth soft edges and it is applied on alpha channel.

1.1.1.2 Multiple Input Single Output Techniques

This section provides an overview of MISO (Multiple Input Single Output) techniques. The input data are multiple static images and the output is an image of higher resolution.

1.1.1.2.1 Interpolation Approach

The easy way for reconstructing an high resolution image from a set of low-resolution images is to use an interpolation-based approach. It allows for reconstructing a high-resolution image by projecting all the acquired low-resolution images to the reference image, then all the information available from each image are fused. Once an HR image is obtained by interpolation, the restoration problem have to be addressed by removing blur and noise.

The super resolution problem cannot be solved well only adopting single image interpolation algorithm. In fact, during the image acquisition process some high-frequency components are irrimediably lost, thus the quality of the resulting image is related to the amount of data available in the image. Three stages are performed succesively in this approach: first the relative motion is calculated, then interpolation on an high resolution grid is performed in order to produce an improved resolution image, finally restoration for blur and noise removal is performed.

Ur and Gross [19] performed a nonuniform interpolation of an ensemble of spatially shifted LR images by utilizing the generalized multichannel sampling theorem of Papoulis [20] and Brown [21]. The interpolation is followed by a deblurring process, and the relative shifts are assumed to be known precisely here. Komatsu et al. [22] proposed a technique for estimating a high resolution image, with reduced aliasing, from a sequence of undersampled frames by applying the Landweber algorithm [23]. In order to measure the relative shift of the cameras the authors proposed a block matching technique. To overcome the problem relating to cameras with the same aperture, they used multiple cameras with different apertures [24]. In [25] Hardie et al. proposed a gradient-based registration algorithm for estimating the shifts between the acquired frames then a weighted nearest-neighbor approach for placing the frames onto a uniform grid to form a final high-resolution image is presented. To reduce effects of blurring and noise caused by the system the authors proposed the application of the Wiener filter, designed using the modulation transfer function (MTF) of the imaging system, to the high-resolution image. In [26] Shah and Zakhor propose a new SR color multiframe algorithm to enhance the spatial resolution of frames in video sequences, using both luminance and chrominance information to estimate the motion field. They also consider the inaccuracy of the registration algorithm by finding a set of candidate motion estimates instead of a single motion vector for each pixel. Nguyen and Milanfar [27] proposed an efficient wavelet-based SR reconstruction algorithm. They exploit the interlacing structure of the sampling grid in SR and derive a computationally efficient wavelet interpolation for interlaced two-dimensional (2-D) data.

The advantage of the non uniform interpolation approach is that it takes relatively low computational load and makes real-time applications possible. However, in this approach, degradation models are limited (they are only applicable when the blur and the noise characteristics are the same for all LR images). Additionally, the optimality of the whole reconstruction algorithm is not guaranteed, since the restoration step ignores the errors that occur in the interpolation stage.

1.1.1.2.2 Frequency Domain Approach

The frequency domain approach makes explicit use of the aliasing that exists in each LR image to reconstruct an HR image. The frequency domain approach is based on the following three principles: i) the shifting property of the Fourier transform, ii) the aliasing relationship between the continuous Fourier transform (CFT) of an original HR image and the discrete Fourier transform (DFT) of observed LR images, iii) and the assumption that an original HR image is band limited.

Tsai and Huang. [1] present an algorithm that improves the resolution of Landsat image data. Landsat acquires several images of partially overlapping areas of the earth in the course of its orbits, thus producing a sequence of similar, but not identical images. Observed images are modeled as under-sampled versions of an unchanging scene undergoing global translational motion. In [28] a frequency domain formulation is proposed, based on the shift and aliasing properties of the continuous and discrete Fourier transforms for the reconstruction of a band-limited image from a set under-sampled, and therefore aliased, observations. Several limitations of the Tsai-Huang method are addressed by Tekalp, Ozkan and Sezan in [29]. The authors propose a frequency domain approach which extends [1] by including the effects of a LSI PSF as well as observation noise. An extension of this approach for a blurred and noisy image was provided Kim, Bose and Valenzuela [30]. They exploit the frequency domain theoretical framework and the global translation observation model proposed in [1] and consider observation noise as well as the effects of spatial blurring.. This method was further refined by Kim and Su [31] to consider different blurs for each LR image. Here, the Tikhonov regularization method is adopted to overcome the ill-posed problem resulting from blur operator.

Periodic sampling is still assumed and a translation-only motion model is used. Theoretical simplicity is a major advantage of the frequency domain approach. That is, the relationship between LR images and the HR image is clearly demonstrated in the frequency domain. The frequency method is also convenient for parallel implementation capable of reducing hardware complexity. However, the observation model is restricted to only global translational motion and LSI blur.

1.1.1.2.3 Spatial Domain Approach

Among spatial domain methods, Keren, Peleg and Brada [32] propose an approach to image registration based on a global translation and rotation model, as well as a two stage approach to super-resolution reconstruction. The first stage is a simple interpolation technique and the second consists in a motion estimation algorithm. An interpolation based technique is proposed by Aizawa, Komatsu and Saito [33]. They examine the problem of acquiring high-resolution imagery from stereo cameras. By considering the possibility of sampling at spatial positions between the array pixels, it is demonstrated that the effective frequency response of the combined (double image) system is increased. In [34] the idea of super-resolution reconstruction from a set of globally translated images of an unchanging 2D scene is considered and compared to a global translation and rotation model used in [32]. A dynamic super-resolution sequence reconstruction from a lower resolution sequence containing sub-pixel shifts is presented in [35]. The main features of this work are related to: the local motion estimation performed using the group delays of local adaptive linear prediction filters, in order to obtain a motion vector for each pixel in the image; and the application of the super-resolution improvement to the sequence images rather than to a prototype image.

Their advantages include a great flexibility in the choice of motion model, motion blur and optical blur, and the sampling process. Another important factor is that the constraints are much easier to formulate.

1.1.1.2.6 Iterative Back-Projection Approach

The first approach to super-resolution based on the iterated process of backprojecting the error between the estimated LR images and the observed data was proposed in [36] and further extended in [37-39]. The algorithm performs an initial estimate of the high resolution image; then, the subsampling/degradation process is simulated in order to deduce the set of LR frames which correspond to the observed input images. The difference (error) between the simulated and the observed frames is computed in order to update the initial HR frame estimate through the error backprojection. The process is iterated in accordance to an error minimization criterion. Only translation and rotation were considered for modeling the HR estimate and LR subsampling.

The relative displacements of the input images at subpixel accuracy are computed and an iterative refinement is adopted to improve accuracy. It is assumed that the imaging process for the observed image sequence (LR) is modeled by:

$$g^{(n)}(\vec{y}) = \sum_{\vec{x}} f^{(n)}(\vec{x}) \cdot h^{PSF}(\vec{x} - \vec{z}_{\vec{y}}\ ) \ (1)$$

where $g^{(n)}$ is the LR image obtained by applying the simulated imaging process to $f^{(n)}$; $f^{(n)}$ is the approximation of f obtained after n iterations; $\vec{x}$ and $\vec{y}$ denote HR and LR pixels respectively, the latter influenced by $\vec{x}$; $\vec{z}_{\vec{y}}$ is the center of the receptive field of $\vec{y}$ in $f^{(n)}$; h^{PSF} is the point spread function of the imaging blur; f is the target HR image to be constructed (unknown).

The iterative update scheme to estimate the HR image f is then:

$$f^{(n+1)}(\vec{x}) = f^{(n)}(\vec{x}) + \sum_{\vec{y} \in U_k Y_{k,\vec{x}}} (g_k(\vec{y}) - g_k^{(n)}(\vec{y})) \cdot \frac{(h^{BP}_{\vec{x}\vec{y}})^2}{c \sum_{\vec{y}' \in U_k Y_{k,\vec{x}}} (h^{BP}_{\vec{x}\vec{y}'})}$$

where $Y_{k,\vec{x}}$ is the set { $\vec{y} \in g_k \mid \vec{y}$ influenced by $\vec{x}$ }; c is a constant normalizing factor and $h^{BP}_{\vec{x}\vec{y}} = h^{BP}(\vec{x} - \vec{z}_{\vec{y}}\ )$.

The error function to be minimized is:

$$\varepsilon^{(n)} = \sqrt{\sum_k \sum_{m_1, m_2} \left(g_k(y_1, y_2) - g_k^{(n)}(y_1, y_2)\right)^2} \qquad (3)$$

Since the choice of the initial estimate does not influence the performance of the algorithm, the average of the LR frames is used as $f_i^{(0)}$; then, it is assumed that $h^{BP} = h^{PSF}$.

The advantage of IBP is that it is understood intuitively and easily. However, this method has no unique solution due to the ill-posed nature of the inverse problem.

1.1.1.3 Multiple Input Multiple Output Techniques

This section provides an overview of MIMO (Multiple Input Multiple Output) techniques. The input data are motion pictures and the output is a video sequence of higher resolution. The SR video approaches reconstruct an image sequence with a higher resolution from a group of adjacent lower-resolution uncompressed image frames or compressed image frames.

1.1.1.3.1 Sliding-window-based SR video approach

The sliding-window-based approach [40–43] is the most commonly-used and direct approach to conduct SR video. The sliding window selects a set of consecutive low-resolution frames for producing one high-resolution image frame; that is, the window is moved across the input frames to produce successive high-resolution frames sequentially. The major drawback of this approach is that the temporal correlations among the consecutively reconstructed high-resolution images are not considered.

1.1.1.3.2 Sequential SR video approach

The major challenge in the SR video problem is how to exploit the temporally correlated information provided by the established high-resolution images and available temporally-correlated low-resolution images respectively to improve the quality of the desired high-resolution images. Elad et al. [44–46] proposed an SR image sequence algorithm based on adaptive filtering theory, which exploits the correlation information among the high-resolution images. However, the information provided by the previously observed low-resolution images is neglected; that is, only a single low-resolution image is used to compute the least-squares estimation for producing one high-resolution image.

1.2. Registration Techniques

Image registration is the process of transforming different sets of data into one coordinate system. It is used to match two or more pictures taken, for example, at different times, from different sensors, or from different viewpoints. Image registration has a lot of applications: matching stereo images to detect object/recognize particular location in navigation system, finding the optimal match for the template in a image and many other. The images need to be aligned each other so that differences can be detected, thus a transformation must be found so that the points in one image can be related to their corresponding points in the other. Image registration algorithms can be classified according to the transformation models they use to relate the target image space to the reference image space. In the following a brief overview of some spacial transformation techniques.

1.2.1 Spatial Transformations

The most common spatial transformations, such as scaling, rotating, skewing, and perspective distortion, are implemented as linear transformation. These kind of transformation can be classified in affine and projective and are usually implemented in a matrix form:

$$\begin{pmatrix} a_1 & a_2 & b_1 \\ a_3 & a_4 & b_2 \\ c_1 & c_2 & 1 \end{pmatrix}$$

where

- $\begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}$ is a rotation matrix

- $\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$ is a translation vector
- $\begin{pmatrix} c_1 & c_2 \end{pmatrix}$ is a projection vector

The transformation between two point is defined by:

$$\begin{pmatrix} a_1 & a_2 & b_1 \\ a_3 & a_4 & b_2 \\ c_1 & c_2 & 1 \end{pmatrix} \times \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix}$$

where x' and y' are the coordinates of the transformed point.

1.2.1.1 Affine Transformations

Affine transformations map straight lines into straight lines, thus it preserves parallelism. The affine transformation is used for scaling, skewing and rotation.

For affine transformations the first two elements in the last line of the transformation matrix mentioned before should be zeros. Affine transformation is also indicated as a transformation of a triangle: since the last row of a matrix is zeroed, three points are enough. The general 2D affine transformation form is:

$$\begin{pmatrix} x_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix} + \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$$

General spatial distortions can occur, such as skew and changes in aspect ratio. The first take into account the distortion of pixel along one or both axis while the second refers to the relative scale between the x and y axes.

The skew components of an affine transformation is represented by:

$$S_x = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad S_y = \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix}$$

This component is represented by:

$$Scale = \begin{pmatrix} s_x & 0 \\ 0 & s_y \end{pmatrix}$$

1.2.1.2 Projective Transformations

The projective transformation shows how the perceived objects change when the view point of the observer changes. It accounts for the distortion which occurs when a 3D scene is projected in a 2D plane. This transformation allows creating perspective distortion. In the projective space, a 3D point is described using a 4-element vector $(X_1, X_2, X_3, X_4)^T$ such that

$$X = X_1 / X_4; Y = X_2 / X_4; Z = X_3 / X_4;$$

where $X_4 \neq 0$. More generally, in n-dimensional space we have:

$$\underbrace{(X_1, X_2, \ldots\ldots\ldots, X_n)^T}_{EuclideanSpace} \rightarrow \underbrace{(\lambda X_1, \lambda X_2, \ldots\ldots\ldots, \lambda X_n, \lambda)^T}_{HomogeneousSpace}$$

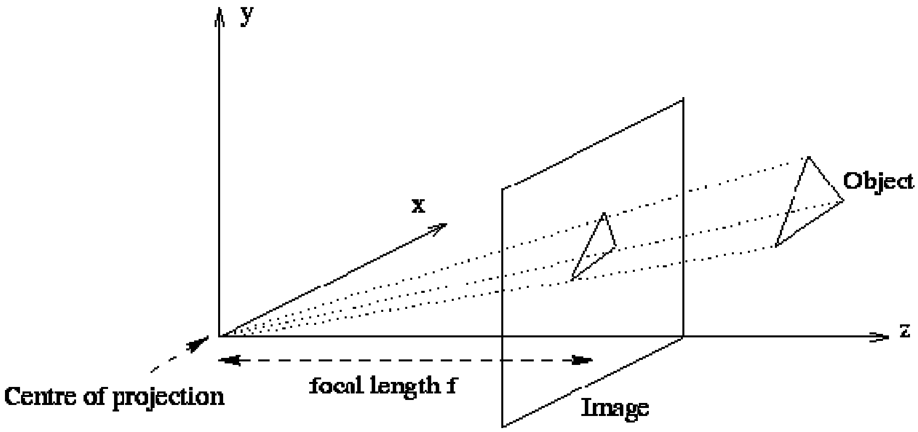
where $\lambda \neq 0$corresponds to the so called homogeneous free scaling parameter.

1.2.1.2.1 The Projection Matrix

To understand how vision might be modeled computationally and replicated on a computer, we need to understand the image acquisition process. The role of the camera in machine vision is analogous to that of the eye in biological systems.

The drop from three-dimensional world to a two-dimensional image is a projection process in which we lose one dimension. The usual way of modeling this process is by central projection in which a ray from a point in space is drawn from a 3D world point through a fixed point in space, the centre of projection. This ray will intersect a specific plane in space chosen as the image plane. The intersection of the ray with the image plane represents the image of the point.

The *pinhole camera* is the simplest, and the ideal, model of camera function. It performs a central projection of point $P = (X,Y,Z)^T$ in the scene onto the plane Z = f, being f the focal distance or distance from projection center to the projection plane.



The mapping is:

$$(X,Y,Z)^T \rightarrow (f\frac{X}{Z}, f\frac{Y}{Z})$$

Dealing with the projection of a 3D point onto an image plane, three different coordinate system are involved: image, world and camera coordinate system.

Let $(X,Y,Z) \in P^3$ be a point in the world coordinate system, $(x,y) \in P^2$ a point in the image coordinate system and (x_c, y_c) a point in the camera coordinate system.

1.2.1.2.2 Camera Coordinate System

The mapping between a 3D point (in the world coordinate system) into a 2D point (in the camera coordinate system) is describe by the following relation:

$$\begin{bmatrix} x_c \\ y_c \\ f \end{bmatrix} = \lambda \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}$$

where $\lambda = f / Z_c$

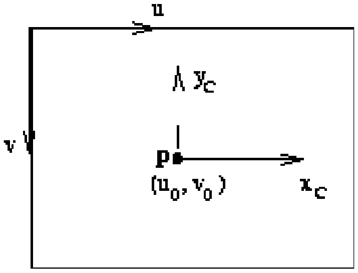
In homogeneous coordinates and up to a scale factor, the previous relation can be written as a linear mapping in the following form:

$$\begin{bmatrix} x_c \\ y_c \\ f \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix}$$

The $3x4$ matrix represents the so called projection matrix which allow for a mapping from a 3D to a 2D point.

1.2.1.2.3 Image Coordinate System

If we consider a point in the image plane, we have the following relationship:



with $k_u x_c = u - u_0$ and $k_v y_c = v_0 - v$ and the units of k are [pixel/length].

The relation between a point in the image plane and a ray in Euclidean 3-space is described by:

$$x_i = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} fk_u & 0 & u_0 \\ 0 & -fk_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ f \end{bmatrix}$$

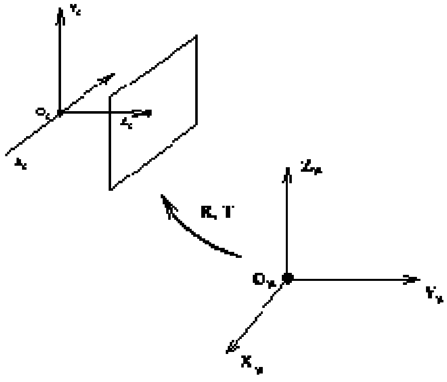
The $3x3$ upper triangular matrix is called the camera calibration matrix and contain the intrinsic camera's parameters.. It can be rewritten as:

$$K = \begin{bmatrix} fk_u & 0 & u_0 \\ 0 & -fk_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \alpha_u & 0 & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

where α_u and α_v are the scaling in the image x and y directions, (u_0, v_0) is the principal point at which the optic axis intersects the image plane. The aspect-ratio is α_v / α_u.

1.2.1.2.4 World Coordinate System

The Euclidean transformation between the camera and world coordinates is generally obtain by a rotation and a translation.



In matrix form we have:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$$

where R is a rotation and T a translation.

Finally, by combining the three matrices we obtain the transformation between a 3D point into a 2D point:

$$x = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = C \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = K[R \mid T] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$$

where $P = [K][R \mid T]$ defines the $3x4$ projection matrix from Euclidean 3-space to an image.

1.3 Rectification Techniques

In a central projection camera model, a three-dimensional point in space is projected onto the image plane by means of straight visual rays from the point in space to the optical centre. Mathematically this process can be described using a $3x4$ projection matrix P, which takes a point in 3-D space in homogeneous coordinates $(X, Y, Z, 1)^T$ and transforms it into a point on the 2-D image plane $(x, y, 1)^T$.

$$\lambda \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = [P] \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \qquad (1)$$

The projection matrix P can be computed from the internal and external camera parameters:

$$P = K \ [R \mid T] \qquad (2)$$

where K is a 3 x 3 upper triangular matrix, called the camera calibration matrix, including the intrinsic camera parameters (focal length, aspect ratio and skew) and

$[R \mid T]$ defines the Euclidean transformation between camera and world coordinates (in general rotations followed by translations), including the external camera parameters, i.e. its position and orientation.

1.3.1 Plane to Plane Homography

In the case where planar surfaces are imaged ($Z = 0$), the transformation is called plane-to-plane homography:

$$\lambda \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = [H] \begin{pmatrix} X \\ Y \\ Z = 0 \\ 1 \end{pmatrix} \qquad (3)$$

The 3x3 transformation matrix, usually called the homography matrix H, has a simpler form than P, but it can be also reduce to:

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} = [K] \; [\mathbf{r}_1 \quad \mathbf{r}_2 \quad \mathbf{t}] \qquad (4)$$

where K is the camera's matrix, $\mathbf{r}_1$ and $\mathbf{r}_2$ are the correspondent columns of the rotation matrix R and $\mathbf{t} = -RC$ with C the camera center.

For this particular case we are dealing with the acquisition of a planar surface. Fig. 1.2 shows the mapping between a 2-D point $x^{'}$ in the object plane $\pi^{'}$ into a 2-D point x in the image plane π.

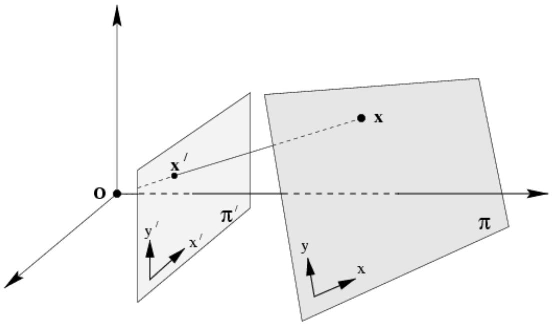


Figure 1.2 Mapping between planes

This process can be described mathematically by a homography matrix H :

$$P_i^{'} = HP_i \qquad (5)$$

where P and $P^{'}$ are 3 x 1 vectors that could correspond to the images of the same points, the former in the plane of the tag and the latter in the plane of the image, while H is the transformation matrix.

If the homography between a plane in the scene and the plane of the image is known, then the image of the planar surface can be rectified into a front-on view. Given four points on the scene plane, with no more than any 2 points collinear, and their corresponding positions in the image (8 equations), H is uniquely determined.

1.3.1.1 The Homography Matrix

The mapping between two planes can be described mathematically by an Homography matrix H :

$$\lambda P_i^{'} = HP_i \qquad H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$$

where P and $P^{'}$ are 3x1 vectors that could correspond to the images of same points, in two different plane, while H is the transformation matrix.

The 3x3 transformation matrix can be easily reduce to:

$$[r_1 \quad r_2 \quad t]$$

where K is the matrix's camera, r_1 and r_2 are the correspondent columns of the rotation matrix R and $t = -RC$ with C the camera center.

1.3.1.2 Scaling factor's computation

There are two methods of dealing with the unknown scale factor λ in a homogeneous matrix:

- choose one of the matrix elements to have a certain value. For example, $h_{33} = 1$
- Solve for the matrix up to scale

If the homography between a plane in the scene and the plane of the image is known, then the image of the planar surface can be rectified into a front-on view. Given four points on the scene plane, with no more than any 2 points collinear, and their corresponding positions in the image (8 equations), H is uniquely determined. Let $P_1'(x'_1, y'_1)$, $P_2'(x'_2, y'_2)$, $P_3'(x'_3, y'_3)$ and $P_4'(x'_4, y'_4)$ be the four corner points of the rectangular object and $P_1(x_1, y_1)$, $P_2(x_2, y_2)$, $P_3(x_3, y_3)$ and $P_4(x_4, y_4)$ their projections obtained using a plane homography transformation, as shown in Fig.1.3.

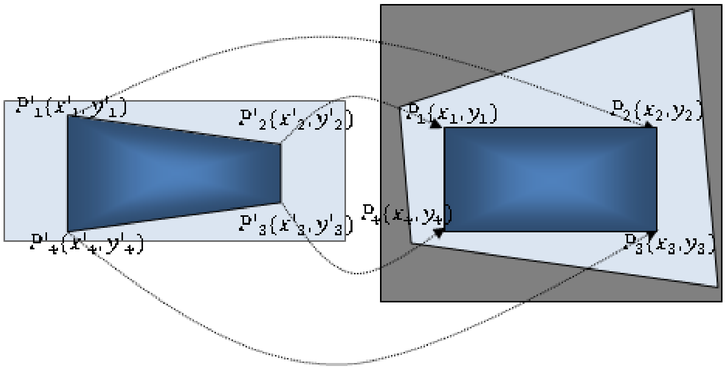


Fig 1.3. The plane homography transformation: on the left the image plane, on the right the rectified frontal view

Corresponding points in two images related by homography are then:

$$x_i' = \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + h_{33}}$$

$$y_i' = \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + h_{33}}$$

In order to remove affine and projective components firstly identify the four corners in the the image, then, using plane homography transformation, map each vertex of the quadrilateral to the corresponding vertex in the known rectangle. Using the previous equations, find the coefficients of the homography matrix H and finally rectify the image to the frontal view. Once the frontal view is recovered from the knowledge of the calculated homography's matrix coefficients, the H matrix can be decomposed by factorization, in its orthogonal $[r_1 \quad r_2 \quad t]$ and upper triangular matrix $[K]$. From the knowledge of the orthogonal matrix the tilt angle φ (rotation around the x-axis), the roll angle ψ (rotation around the y-axis), the pan angle θ (rotation around the z-axis) and the translation along the three axis can be determined, thus the orientation and position of camera in the scene. From $[K]$, given the focal length of the camera, the other internal camera parameters can be deduced.

1.4 Image Analysis

The task of finding similarity correspondences between two images of the same scene or object has a great importance. Image analysis refers to the extrapolation of meaningful information from images, called feature, in order to simplify the amount of resources required to describe a large set of data accurately. A feature is defined as an "interesting" part of an image, and represents the starting point for every image analysis algorithm.

In the following a brief overview of the most important feature detectors: Scale Invariant Feature Transform (SIFT) and Speeded up Robust Features (SURF) will be presented.

1.4.1 Feature Detection Description and Matching

SIFT and SURF are the two most famous algorithms for feature detection and description. They are able to detect and describe local features in images. For any object in an image, interesting points on the object can be extracted to provide a "feature description" of the object. This description, extracted from a training image, can then be used to identify the object when attempting to locate the object in a test image containing many other objects. To perform reliable recognition, it is important that the features extracted from the training image be detectable even under changes in image scale, noise and illumination.

1.4.1.1 SIFT

The SIFT descriptor [47] provide a set of features of an object that are not affected by many of the image transformations like image rotation, scale illumination and are shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination.

The cost of extracting these features is minimized by taking a cascade filtering approach, in which the more expensive operations are applied only at locations that pass an initial test. The algorithm consist of four step:

Scale-space extrema detection

The first stage of computation searches over all scales and image locations. This can be efficiently achieved using a "scale space" function based on the Gaussian function, that is a convolution between a variable-scale Gaussian $G(x, y, \sigma)$ and the input image $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$

In order to detect stable keypoint locations in the scale-space, difference of Gaussians is used. It can locate scale-space extrema by computing the difference between two images, one with scale k times the other. $D(x, y, \sigma)$ is then given by:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma)$$

To detect the local maxima and minima of $D(x, y, \sigma)$ each point is compared with its 8 neighbors at the same scale, and its 9 neighbors up and down one scale. If this value is the minimum or maximum of all these points then this point is an extrema.

Keypoint localization

At each candidate location, a detailed model is fit to determine location and scale. Keypoints are selected based on measures of their stability. This is achieved by calculating the Laplacian value for each keypoint found in the previous step. The location of extremum, z, is given by:

$$z = -\frac{\partial^2 D^{-1}}{\partial x^2}\frac{\partial D}{\partial x}$$

If the function value at z is below a threshold value then this point is excluded.

Orientation assignment

One or more orientations are assigned to each keypoint location based on local image gradient directions. All future operations are performed on image data that has been transformed relative to the assigned orientation, scale, and location for each feature, thereby providing invariance to these transformations.

In order to find an orientation:

- Use the keypoints scale to select the Gaussian smoothed image L, from above
- Compute gradient magnitude:

  $m(x, y) = \sqrt{(L(x+1, y) - L(x+1, y))^2 + (L(x, y+1) - L(x, y-1))^2}$
- Compute orientation

  $\theta(x, y) = \tan^{-1}(((L(x, y+1) - L(x, y-1)) / (L(x+1, y) - L(x+1, y)))$
- Form an orientation histogram from gradient orientations of sample points
- Locate the highest peak in the histogram. Use this peak and any other local peak within 80% of the height of this peak to create a keypoint with that orientation
- Some points will be assigned multiple orientations
- Fit a parabola to the 3 histogram values closest to each peak to interpolate the peaks position

Keypoint descriptor

The local image gradients are measured at the selected scale in the region around each keypoint and used to create keypoint descriptors. The contribution of each gradient orientation in its histogram is weighted by its gradient magnitude and by a Gaussian weighting function with σ equal to one half the width of the normalized image patch. Keypoint descriptors typically uses a set of 16 histograms, aligned in a 4x4 grid, each with 8 orientation, thus, the resulting descriptor is of dimension 128.

1.4.1.2 SURF

SURF [48,49] is an efficient scale and rotation invariant interest point detector and descriptor. It allows for quick and effective feature detection even against different image transformations like image rotation, scale illumination and small viewpoint changes.

Much of the performance increase can be attributed to the use of an intermediate image representation, known as the Integral Image that can be rapidly computed from an input image [50]. In the following a brief summary of its construction process.

Interest point detection

SURF is a Hessian matrix based interest point detector. It searches for blob-like structure at locations where the determinant of this matrix is maximal. Given a point $X = (x, y)$ in an image $I(x, y)$, the Hessian matrix $H = (X,\sigma)$, as function of both space X and scale σ, is defined as follows:

$$H(X,\sigma) = \begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix}, \qquad (4)$$

where $L_{xx}(X,\sigma)$ refers to the convolution of the second order Gaussian derivative $\frac{\partial^2 g(\sigma)}{\partial x^2}$ with the image at point $X = (x, y)$ and similarly for $L_{yy}(X,\sigma)$ and $L_{xy}(X,\sigma)$. These derivatives are known as Laplacian of Gaussians. The approximated determinant of the Hessian represents the blob responses at location $X = (x, y)$ in the image. In order to detect interest points over different scale a non maxima suppression in a 3 x 3 x 3 neighbourhood is applied. To do this each pixel in the scale-space is compared to its 26 neighbours, comprised of the 8 points in the native scale and the 9 in each of the scales above and below. Finally the maxima of the determinant of the Hessian matrix are then interpolated in both space and scale to sub-pixel accuracy.

Interest point descriptor

The SURF descriptor describes the distribution of pixel intensities within a scale dependent neighbourhood of each interest point detected by the Fast-Hessian. Integral images in conjunction with Haar wavelets are used in order to increase robustness and decrease computation time. Haar wavelets are used to find gradients in the x and y directions. The first step in descriptor's extraction consists of fixing a reproducible orientation based on information from a circular region around the interest point. Then, a scale dependent window aligned to the selected orientation is constructed and a 64-dimensional vector (SURF descriptor) is extracted from it. The dominant orientation is estimated by calculating the sum of all responses within a circle segment covering an angle of $\pi/3$ around the origin. At each position, the two summed x and y responses are used to form a new vector.

The longest vector defines the orientation of the interest point. The first step for the extraction of the descriptor is to construct a square region aligned with the selected orientation around the interest point. It contains the pixels which will form entries in the descriptor vector and is of size 20σ, where σ refers to the detected scale. A further division into $4x4$ regular sub regions is performed within each Haar wavelets of size 2σ, calculated for $5x5$ regularly spaced sample points. Hence, each sub-region has a four dimensional descriptor vector, thus concatenating this for all $4\,x4$ sub-regions a descriptor vector of length 64, invariant to different image transformation is obtained.

Chapter 2 Client-Side Super Resolution Image Presentation

2.1 Introduction

Latest years have been characterized by the surprising success of distributed and collaboratively created multimedia databases over the web for a great variety of purposes: from scientific to commercial, from educational to entertainment. The end-users are taking more and more an active part in content generation, which is the main cause of the impressive growth of the multimedia data volume. Not only do they collaborate to the generation of the content, but they also contribute to the storage and distribution, exploiting the great potentials of peer-to-peer networks. To enable an efficient and fast browsing of such huge amounts of multimedia data, it is important to develop tools for analyzing and describing the content, handle queries from the end-users, and provide the results. Accordingly, a procedure for data analysis, indexing and presentation has become a requirement for efficient content management and search. It mainly requires the accomplishment of the following tasks: feature extraction, structure analysis, abstraction and indexing. The first task is aimed at providing the major characteristics of the multimedia data (such as color, texture, shape, structure, layout, and motion) that can be converted into semantic concepts. Data structure parsing is the next step in overall multimedia-content analysis and is the process of extracting spatial and temporal structural information. Multimedia data abstraction is the process of creating a glance of the multimedia information, such as sub-sampled version in case of still picture browsing. Based on the output of the previous tasks, video indices are built so as to enable a fast browsing of the visual content. The quality of experience in database browsing is then enhanced by integrating such features and tools at the server side with appropriate services at the client side that improve the presentation of the multimedia data. For instance, spatial details are predicted from low-resolution images to obtain super-resolution images with respect to the provided thumbnails.

Within the described context, this chapter addresses the problem of multimedia data presentation by proposing a client-side super-resolution approach. The considered scenario is that of a user browsing the information in a huge database of images transmitted at low spatial resolution (thumbnails), which are then enhanced by increasing the image resolution with adaptive image interpolation. The aim is to provide the user with additional details that improve the quality of database browsing requiring no additional transmission overhead. In the proposed approach, the low-definition image is first analyzed in order to identify several features that are significant for visual rendering and scene understanding. Such segmentation is based on local frequency composition: uniform regions, edges and textures. The identified regions are then treated differently depending on the relative visual significance. Each region is further analyzed and a different interpolation approach is adopted, ranging from plain linear interpolation for homogeneous areas to edge area analysis and selective anisotropic interpolation.

2.2 Image Super Resolution with adaptive interpolation

The proposed technique pursues the ambition of making the image subsampling process visually reversible by reconstructing a high-definition image from a single low-resolution sample. The process is based on region-of-interest (ROI) segmentation. The LR image is first analyzed in order to identify the features that are significant for visual rendering and scene understanding. Such segmentation is based on local frequency content and is used to discriminate between three principal signal behaviors: smooth regions, textures, and edges with relevant surrounding areas (Fig. 1).

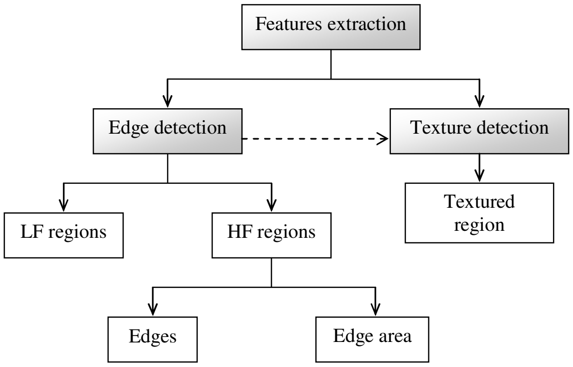


Fig. 2.1. Exploited features (LF = Low frequency, HF = High frequency)

The identified regions are then treated differently depending on the their visual significance. In particular, an adaptive interpolation approach is adopted, ranging from plain linear interpolation for homogeneous areas to edge area analysis and selective anisotropic interpolation.

In the following paragraphs, the proposed algorithm is described is terms of features extraction, needed for determining the regions of interest, and region processing.

The feature extraction and region processing procedure is performed on the HR scale. A bilinear oversampled version of the LR image is assumed as starting model. In this way, while the feature-map is defined at full-resolution, all processing is carried out based on currently known or reconstructed pixels.

2.1.1 Feature extraction

Feature extraction is performed to identify the regions of interest and to produce the corresponding ROI map. It must be noted that the synthetic information produced during this process, such as background extension or the quantity and characteristics of edges and textures can be easily used for low-level image classification. In the context of content-based image indexing the low-level features represent a useful output for further integration, which is not addressed in this work.

2.1.1.1 Edges

Edges comprise a fundamental role in scene representation and understanding. Their correct reconstruction is then essential for reliable high-definition reproduction.

Edge detection is performed through a Canny-like filter processing. The first Gaussian derivative is computed both in magnitude and phase to determine local maxima in each edge direction. Although the proposed work relies on a simplified implementation that does not guarantee single edge detection, the resulting edge maps proved to be accurate enough for the following processing.

2.1.1.2 Edge area

The areas surrounding the edges are critical for the rescaling operators as much as the edge themselves. In fact, since edges represent an abrupt change in signal amplitude, two conditions are likely to happen:

1. the surrounding areas presents a considerable gradient in the edge normal direction and/or
2. the average signal behavior differs significantly between the areas surrounding the edge in the opposite sides

The edge area is simply obtained as the difference between the edge map and its morphological dilation with circular structural elements. Further analysis is performed during the feature processing phase (Section 2.1.2.3).

2.1.1.3 Textures

In order to extract relevant textural features, two conditions are considered:

- textured areas should have a statistically relevant high-frequency content and
- they should show some structure regularity.

Such characteristics allow us to distinguish between plain edges, which are locally isolated and don’t present any significant specific structure, and complex edge patterns that generally indicate statistical textures. The proposed approach suits well with the given general-purpose requirement.

Sobel magnitude and phase are first considered for texture segmentation. Through an overlapping block analysis, the texture requirements are translated in terms of average edge magnitude and direction.

The Sobel magnitude is first thresholded in order to reduce noise and weak edges contribution. A candidate block is further considered if its average edge magnitude is greater than half of the maximum signal strength. The resulting edge map is then analyzed for structure. The Sobel phase is quantized to 8 principal directions, as shown in Table I.

Table I. Quantization of the Sobel phase.

Θ	ΔΘ	symb.	Θ	ΔΘ	symb.
0	337.5 : 22.5	30	180	157.5 : 202.5	150
45	22.5 : 67.5	60	225	202.5 : 247.5	180
90	67.5 : 112.5	90	270	247.5 : 292.5	210
135	112.5 : 157.5	120	315	292.5 337.5	240

The histogram of the candidate block is then produced to analyze its structural content (Fig. 2.2). Given the eight quantized directions, a probability threshold of 0.125 is set to look for relevant phase components. If the block is found to include at least two opposite directions greater than the designed threshold, then the corresponding neighborhood is tagged as textured area.

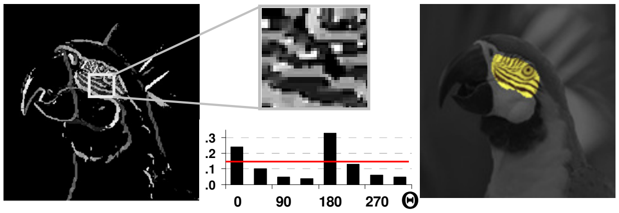


Fig. 2.2 Texture analysis from Sobel phase. From left: quantized Sobel phase, detail and block histogram, result of final segmentation.

Selectivity may be increased by limiting the number of allowed peaks in the direction distribution to 2/3 or redefining the direction quantization by narrowing the fan angle.

2.1.1.4 Background

The background area is simply what is left from all other processing.

2.1.2 Region processing

Region processing is performed in inverse order with respect to feature extraction. In this way, the proposed method first attempts to solve the approximation of simple regions and then to serially recover crucial missing information.

2.1.2.1 Background

Background reconstruction is performed through simple bilinear interpolation. Such a choice represents a good tradeoff between performance and low computational complexity in regions that present no particular challenges. In fact, background pixels have very small high-frequency content and the addition of slowly-varying artificial gradients does not decrease the visual quality. It has to be noted that, on average, background pixels are the most numerous, so that a good portion of the high-resolution signal is reconstructed in this first step very easily.

Linear pixel reconstruction is performed with a block-based scan of the low-resolution image and the feature-map in parallel (Fig. 2.3).

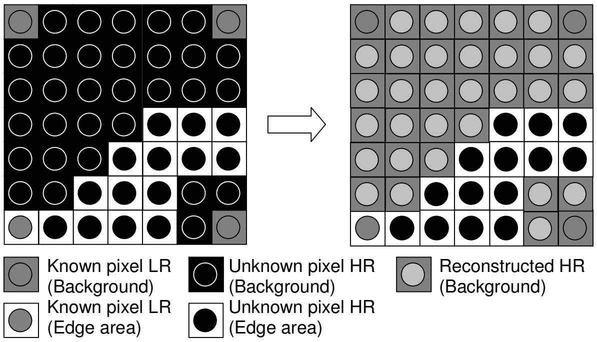

Known pixel LR (Background)
Unknown pixel HR (Background)
Reconstructed HR (Background)
Known pixel LR (Edge area)
Unknown pixel HR (Edge area)

Fig. 2.3. Background interpolation

2.1.2.2 Textures

Textured areas represent regions with relevant high frequency content arranged in a structured pattern which cannot be treated either with bilinear interpolation or contour-based reconstruction techniques. In fact, bilinear interpolation does not perform properly with high-frequency components, introducing strong blurring artifacts. Besides, edge preserving approaches, such as the one described in the following section, do not cope well with complex edge structures. Then, we propose the use of bicubic interpolation to approximate these regions. Such choice suits the general-purpose requirement, since it does not address any specific textural structure, and provides better results than the bilinear interpolation with reduced jagged artifacts.

2.1.2.3 Edge area

The edge neighborhoods are reconstructed through local anisotropic approximation. At each iteration, the edge map is analyzed for connected component labeling in order to distinguish between regions belonging to either side of each edge, referred to as A and B, as shown in Fig. 2.4.

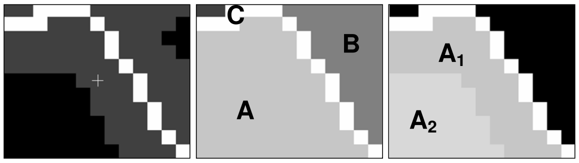


Fig. 2.4. Segmentation and edge areas labeling. From left: original feature-map, result of labeling, region weighting for the considered area

When processing a pixel belonging to region A, only A neighbors are considered, so that statistics from the other side of the edge do no influence the local restoration. With reference to Fig. 2.4, A pixels are split into two classes: A_1, A pixels belonging to the edge area, and A_2, A pixels belonging to the nearby background that are also considered for the local reconstruction. The interpolation is based on bilateral filtering [51], which relies on dynamic FIR (finite impulse response) kernels built from known pixels through three weighting contributions:

Spatial distance: $W_S^{i,j}(h,k) = \exp\left(-\frac{d^2([i,j],[i-h,j-k])}{2\sigma_S^2}\right)$ (1)

Amplitude distance: $W_R^{i,j}(h,k) = \exp\left(-\frac{(\hat{I} - I(i-h,j-k))^2}{2\sigma_R^2}\right)$ (2)

Area of interest :

$$W_A^{i,j}(h,k) = \begin{Bmatrix} \alpha_1 & p_{h,k} \in A_1 \\ \alpha_2 & p_{h,k} \in A_2 \\ 0 & otherwise \end{Bmatrix} \quad (3)$$

In (1), (i,j) is the kernel center in respect of the original image I, $d(x,y)$ is the Euclidean distance function and σ_S^2 is the spatial variance. In (2), $\hat{I}$ is the average signal amplitude in the surrounding of the kernel center, $I(x,y)$ is the signal amplitude at $p_{x,y}$ and σ_R^2 is the amplitude variance. The coefficients α_1 and α_2 are inversely proportional to the number of pixels belonging to the smooth and textured regions whose values have already been estimated. Finally, each kernel coefficient is computed as:

$$W^{i,j}(h,k) = \frac{W_S^{i,j}(h,k) \cdot W_R^{i,j}(h,k) \cdot W_A^{i,j}(h,k)}{\sum\limits_{h,k \in \mathrm{ker}nel} W_S^{i,j}(h,k) \cdot W_R^{i,j}(h,k) \cdot W_A^{i,j}(h,k)} \qquad (4)$$

Fig. 2.5 illustrates the three weighting contributions and the final kernel, *W*.

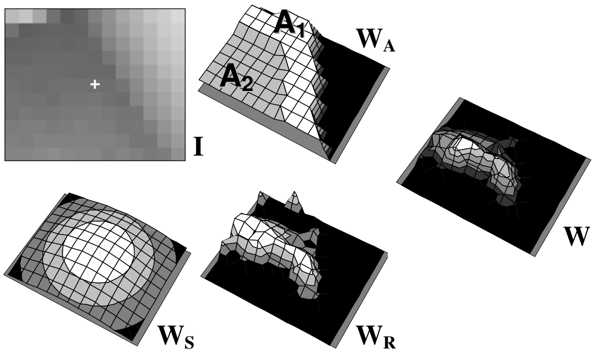


Fig. 2.5. Edge area pixel reconstruction.

2.1.2.4 Edges

Edge pixels are reconstructed through median filtering of a small neighborhood of known pixels. Such solution combines low computational cost with the preservation of edge sharpness.

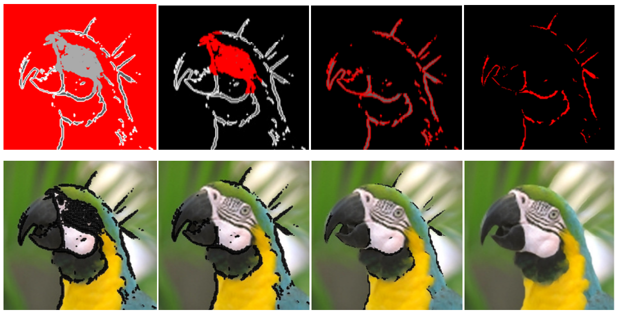


Fig. 2.6. Complete processing; above: feature-map, below: reconstructed image; from left: background, textured area, edge area and edges.

2.3 Experimental Results

The proposed method has been evaluated with a test set of 20 images @24bpp, chosen among the Kodak [52] and Canon [53] databases and other classical image processing test sets [54]. The test images have been selected with the purpose of presenting a broad range of signal behaviors, in terms of high frequency content and textures.

To provide objective results, the testing procedure consisted in preliminary subsampling the original image at a given zoom factor and reconstructing the signal with several interpolative methods. Then, PSNR was computed between the original and the reconstructed signal (Fig. 2.7). Subsampling is executed through block average. The proposed method (SR) is compared with the nearest neighbor (NN), bilinear (BL) bicubic (BC) and bilateral filter (BF) interpolation. Results are expressed in terms of average PSNR (Fig. 2.8) and PSNR standard deviation (Fig. 2.9) among the complete test set.

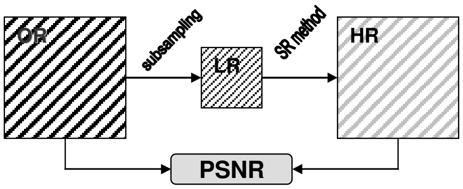


Fig. 2.7. Evaluation of the objective results.

As expected, reconstruction quality decreases significantly as the zoom factor increases. Such behaviour characterizes all methods and derives from the increasing lack of information. In fact, the subsampled image constitutes the only piece of information for all reconstruction methods.

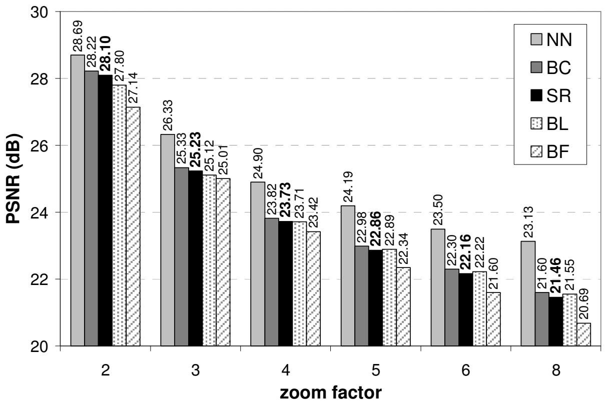


Fig. 2.8. Objective results: average PSNR

Objective results show that the proposed method results lies between the bicubic and the bilinear interpolation, whereas the bilateral filtering interpolation provides the worst results. The apparently unpredicted performance of the nearest neighbour interpolation is easily explained. Since subsampling is carried out through block averaging, nearest neighbour substitution simply assigns the local average to unknown pixels, thus approximating their value with the best esteem in terms of mean square error, thus PSNR.

From the previous considerations, PSNR is apparently inadequate in providing a reliable quality index. It is a measure that provides only a rough performance indication and cannot be considered as an accurate indication of reconstruction quality. An interesting utilization of the PSNR is the computation of the standard deviation, which is always lower for the proposed method than for bilinear or bicubic interpolation. Such index gives us a measure of dispersion, thus unreliability of the method.

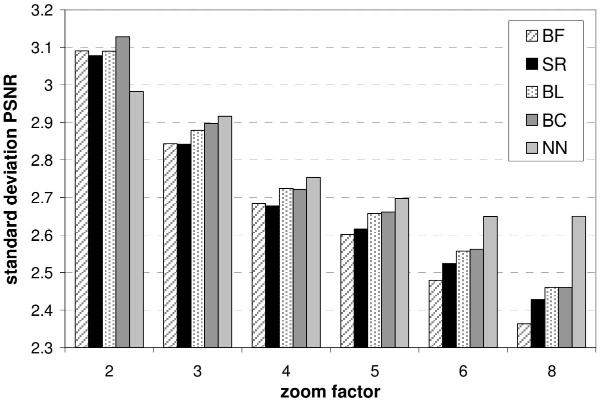


Fig. 2.9. Objective results: PSNR standard deviation

Average SSIM [55] results are also provided in Fig. 2.10 for three zoom factors (2, 4 and 8). The index measures the comparative image quality in a range 0÷1 based on the degradation of structural information, providing a better visual quality esteem than PSNR. In this case, SR outperforms BL and is very close to BC. However, since NN still prevails on other methods, SSIM cannot still be taken as a perfect indication.

Visual results are provided in Figs. 2.11, 2.12 and 2.13 in order to subjectively evaluate the proposed method. The original image detail (OR) is compared with the proposed method and all those used in the objective evaluation. The proposed method results appear visually more pleasant than the competitors.

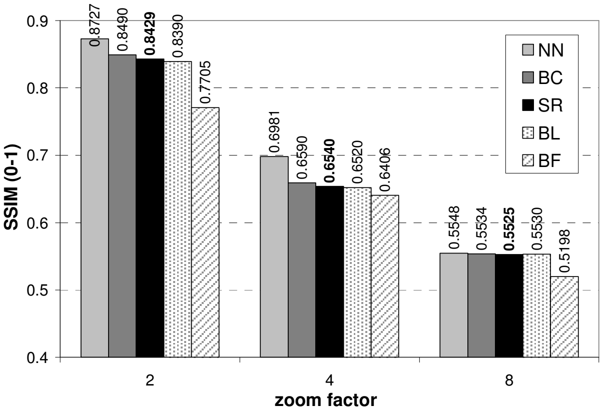


Fig. 2.10. Objective results: average SSIM

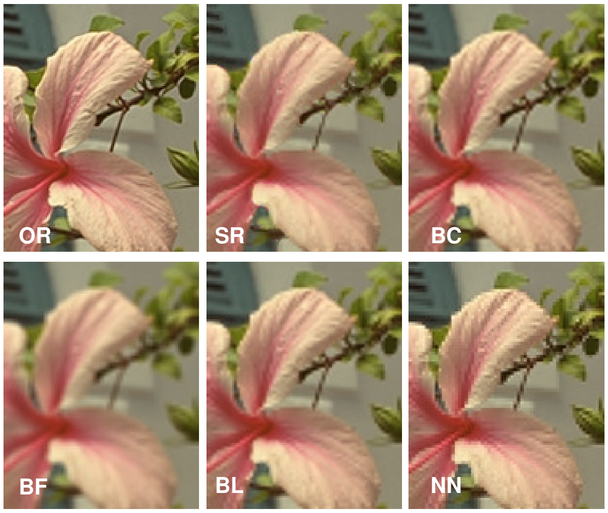


Fig. 2.11. Visual results: “kodim07” at 2× zoom factor

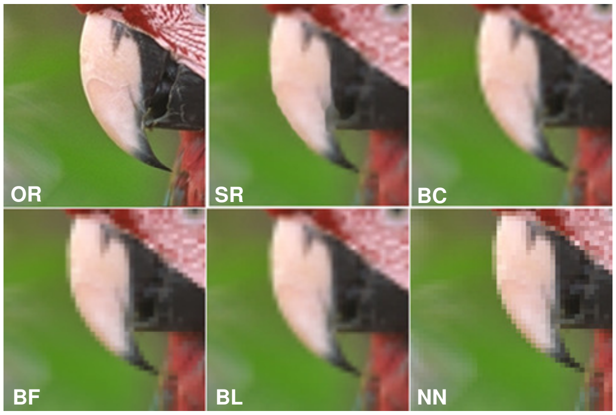


Fig. 2.12. Visual results: “kodim23” at 4× zoom factor

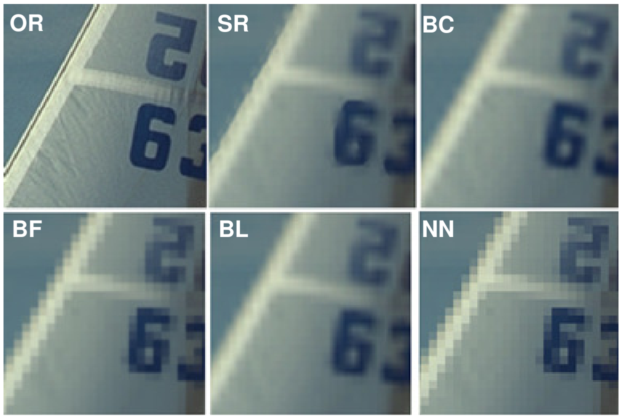


Fig. 2.13. Visual results: “kodim10” at 6× zoom factor

2.4 Conclusions

In this work we presented an adaptive technique aimed at reconstructing high-definition images from low-resolution data, as resulting from sub-sampling or thumbnailing. The proposed algorithm is based on a composite interpolative approach. A feature-map is first computed from the original image. Features include edges, edge area, texture and background. Different interpolation methods are then applied to the recognized areas. Reconstruction methods range from bilinear to ad-hoc anisotropic interpolation. Results are promising, especially when considering visual quality rather than PSNR-based objective evaluation. Future developments comprise the assimilation of edge features to edge areas and further customization for textured regions.

Chapter 3 Video Super-Resolution Reconstruction

3.1 Introduction

On one hand, latest years have been characterized by an incredible proliferation and surprising success of user generated multimedia contents, distributed and collaborative multimedia database over the web. This brought to serious issues related to their management and maintenance: bandwidth limitation and service costs are important factors when dealing with mobile multimedia contents' fruition. On the other hand, the current multimedia consumer market is characterized by the advent of cheap but rather high-quality high definition displays. However, this trend is only partially supported by the deployment of high-resolution multimedia services, thus the resulting disparity between content and display formats have to be addressed and older productions need to be either re-mastered or post-processed in order to be broadcasted for HD exploitation. In the presented scenario, super-resolution reconstruction represents a major solution. Image or video super resolution techniques allow for restoring the original spatial resolution from low-resolution compressed data. In this way, both content and service providers, not to tell the final users, are relieved from the burden of providing and supporting large multimedia data transfer.

3.2 Video Super Resolution based on back projection and motion estimation

Bandwidth limitation and service costs are important factors when dealing with mobile multimedia contents' fruition. Super-resolution reconstruction might be a relevant solution, since it allows for restoring the original spatial resolution from low-resolution compressed data. In this way, both content and service providers, not to tell the final users, are relieved from the burden of providing and supporting large multimedia data transfer. Nowadays, the incredible production of user-generated multimedia contents is leading to serious issues related to their management and maintenance. The combination of increasing bandwidth availability and the development of software technologies allowing for the distributed and collaborative creation of multimedia objects has led to the proliferation of user-generated video communities and, more generally, multimedia information sharing. The massive production and distribution of multimedia data is already a relevant issue for service providers, device designers and software developers: the former are requested to satisfy the ever-growing bandwidth demand; device designers must face the challenge of developing more powerful and compact devices; the latter have to provide better applications and programming frameworks. A fourth category comprises image processing and compression standards professionals, who try to develop better algorithms for coding and reconstructing/recovering the multimedia signals.

Given such scenario, the archetypal use case is that of a user browsing through a huge video database. In order to minimize the bandwidth requirement and the latency, the video streams should be efficiently coded and transmitted at low spatial resolution. Then, the idea is to increase the video stream resolution through super-resolution reconstruction in order to provide the user with additional details that enhance the quality of multimedia browsing, preventing the transmission of additional overhead.

Within the devised context, this chapter addresses the problem of super-resolution restoration of video sequences by proposing an approach based on back projection and motion estimation. The resolution enhancement is performed from multiple under-sampled and degraded frames by taking advantage of the additional spatio-temporal data available in the image sequence. In particular, the motion of both scene and camera is the cause for contiguous frames containing similar, but not identical information. The reconstruction of visually superior frames at higher resolution is then based on the exploitation of such inter-frame information.

Given the observer's motion, each frame shows further details if compared to adjacent frames. Then, resolution enhancement can be achieved by identifying the corresponding image portions through motion estimation and combining the information from a limited number of frames. Although the provided example constitutes an ideal case since the observer's motion results in the natural zooming of the scene, similar considerations are also possible when dealing with different motion models and sub-pixel reconstruction.

The proposed technique is aimed at reconstructing a high-definition video from a limited number of frames extracted from a low-resolution sequence, without any preliminary knowledge of the high-definition data. The process is based on backprojection and motion estimation. For any given frame, a sliding window determines the set of low resolution frames to be processed in order to produce the output stream. The window is shifted forward to produce successive super-resolution frames of the output sequence, as shown in Fig. 3.1

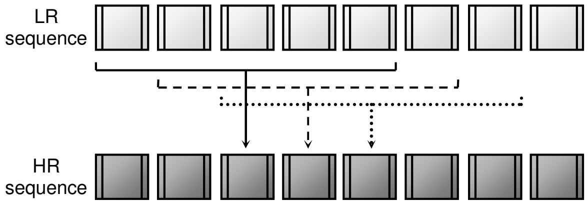


Fig. 3.1 Super-resolution video enhancement from a LR image sequence.

The main idea is that each pixel in a LR frame is a "projection" of a region in the scene. The HR image is constructed using an approach similar to the back projection method used in CAT (Computed Aided Tomography). Accurate knowledge of the relative scene locations sensed by each pixel in the observed images (LR) is necessary for super-resolution. This information is available in image regions where local deformation can be described by some parametric functions.

In the following paragraphs, the proposed algorithm is described is terms of imaging process, super-resolution and motion estimation in order to achieve the reconstruction of higher resolution frame which approximates the original one as accurately as possible.

3.2.1. Background Theory

The first approach to super-resolution based on the iterated process of backprojecting the error between the estimated LR images and the observed data was proposed in [37] and further extended in [38-39]. The algorithm performs an initial estimate of the high resolution image; then, the subsampling/degradation process is simulated in order to deduce the set of LR frames which correspond to the observed input images. The difference (error) between the simulated and the observed frames is computed in order to update the initial HR frame estimate through the error backprojection. The process is iterated in accordance to an error minimization criterion. Only translation and rotation were considered for modeling the HR estimate and LR subsampling.

The relative displacements of the input images at subpixel accuracy are computed and an iterative refinement is adopted to improve accuracy. It is assumed that the imaging process for the observed image sequence (LR) is modeled by:

$$g^{(n)}(\vec{y}) = \sum_{\vec{x}} f^{(n)}(\vec{x}) \cdot h^{PSF}(\vec{x} - \vec{z}_{\vec{y}}) \qquad (1)$$

where $g^{(n)}$ is the LR image obtained by applying the simulated imaging process to $f^{(n)}$; $f^{(n)}$ is the approximation of f obtained after n iterations; $\vec{x}$ and $\vec{y}$ denote HR and LR pixels respectively, the latter influenced by $\vec{x}$; $\vec{z}_{\vec{y}}$ is the center of the receptive field of $\vec{y}$ in $f^{(n)}$; h^{PSF} is the point spread function of the imaging blur; f is the target HR image to be constructed (unknown).

The iterative update scheme to estimate the HR image f is then:

$$f^{(n+1)}(\vec{x}) = f^{(n)}(\vec{x}) + \sum_{\vec{y}\in U_k Y_{k,\vec{x}}} (g_k(\vec{y}) - g_k^{(n)}(\vec{y})) \cdot \frac{(h^{BP}_{\vec{x}\vec{y}})^2}{c \sum_{\vec{y}'\in U_k Y_{k,\vec{x}}} (h^{BP}_{\vec{x}\vec{y}'})} \qquad (2)$$

where $Y_{k,\vec{x}}$ is the set $\{ \vec{y} \in g_k \mid \vec{y}$ influenced by $\vec{x} \}$; c is a constant normalizing factor and $h^{BP}_{\vec{x}\vec{y}} = h^{BP}(\vec{x} - \vec{z}_{\vec{y}})$.

The error function to be minimized is:

$$\varepsilon^{(n)}=\sqrt{\sum_{k}\sum_{m_1,m_2}\left(g_k(y_1,y_2)-g_k^{(n)}(y_1,y_2)\right)^2} \qquad (3)$$

Since the choice of the initial estimate does not influence the performance of the algorithm, the average of the LR frames is used as $f_i^{(0)}$; then, it is assumed that $h^{BP}=h^{PSF}$.

3.2.2. Super-Resolution Approach

Starting from the devised scheme, the proposed work introduces several changes in order to outperform alternative techniques.

Let f_i denote the target frame to be reconstructed through the super-resolution method; we then extract k frames from the original LR video sequence: $(k-1)/2$ past and $(k-1)/2$ future frames. Differently from Peleg and Irani's method, the initial estimate for the high-resolution frame ($n=0$) is a linear interpolation of the low-resolution one, g_i. The blocks of the neighboring LR frames which are found to be significantly similar to those of the reference frame are merged into the HR approximation. Such process resembles a projection, and is done according to the zoom factor and the estimated motion, in order to reconstruct the high-definition data. The block-based motion estimation is described in detail in Section 3.2.3. Residual information is restored through linear interpolation.

Such process is repeated for each of the k LR frames in order to obtain an approximation of k HR frames. They are then subsampled with the h^{PSF} filter to obtain the simulated LR frame sequence $g_{i-((k-1)/2)}^{(n)},\ldots,g_i^{(n)},\ldots,g_{i+((k-1)/2)}^{(n)}$. The difference between simulated and reconstructed LR frames are computed and the error is backprojected into the HR estimate in order to refine the restoration process.

The procedure is iterated until the error becomes appreciably small or a maximum number of iterations, n, is reached. Finally the reconstructed frame, $f_i^{(n)}$, is assumed as high-resolution approximation of g_i, $f_i^{(n)} \cong f_i$.

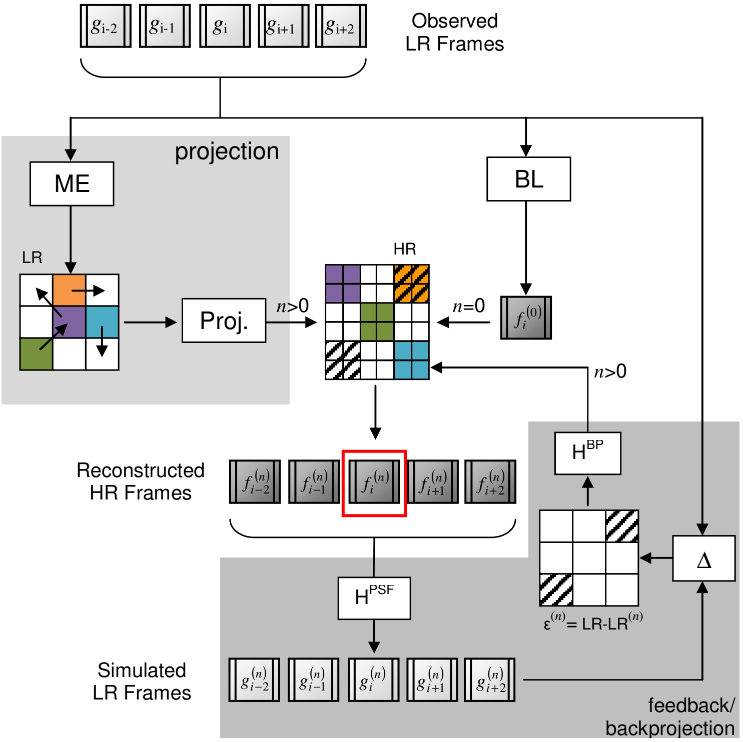


Fig. 3.2. Block scheme of the proposed method.

3.2.3. Motion Estimation

Motion estimation plays an important role in the video super-resolution reconstruction. Among motion estimation techniques, [56] and [57] have been considered, in which a three-step search algorithm is proposed. It employs a center-biased checking point pattern in the first step, which is derived by making the search adaptive to the motion vector distribution, and a halfway-stop technique to reduce the computational cost. In details, by considering a block of size $N \times N$, the block motion estimation searches for a motion vector in a previous frame that yields the minimum block distortion measurement (BDM) in the scanning area. To do this, a multiple stage search is implemented:

1. the central point, the 8 points at $p/2$ in the scanning area ($p \times p$) and 8 extra neighbors points at 3-pixel distance are checked;
2. a halfway stop technique is used to estimate the stationary or quasistationary block's motion:
   a. if the minimum BDM in step 1 occurs at the search windows center, stop the search (first step stop);
   b. if the minimum BDM point in step 1 is one of the 8 neighbors of the window center, the search is performed for the 8 neighboring points of the minimum only (second step stop).

A complete three step search is only performed when the minimum BDM point at the first step is not the window center, nor any of its 8 neighbors.

3.2.4. Method's parameters

In this section an overview of the method's main parameters is given. In particular:

- $f_i^{(0)}$ is the initial HR frame estimate. It is computed as linear interpolation of the corresponding LR frame only, g_i.
- h^{PSF} is the point spread function of the imaging system. In this implementation it represents a Gaussian filter.
- For each HR frame, $k = 5$ LR frames are considered for processing.
- A number of $n = 5$ maximum iterations is imposed.
- The zoom factor tested are $4\times$ and $8\times$. The blocks are 4×4 and 8×8 pixel wide respectively.
- The scanning area is 16×16 pixel.
- The mean square error is considered for BDM.

3.2.5Experimental Results

The proposed method has been evaluated with a test set of 7 video sequences in the 4:2:0 YUV format, chosen among classical video processing test sets [58]. The test video sequences have been selected with the purpose of presenting a broad range of signal behaviors, in terms of different motion.

To provide objective results, a subsampled video sequence is preliminarily produced (LR) from the original video (OR) and used as input sequence for the devised algorithm at any given zoom factor. Then, PSNR is computed between the original and the reconstructed (HR) signal (Fig. 3.3).

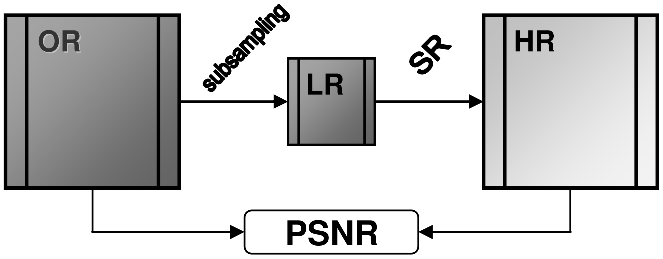


Fig. 3.3. Evaluation of the objective results.

The proposed method (SR) is computed for two different parameters sets (SR_A: $\sigma = 10; c = 0.1$, SR_B: $\sigma = 2; c = 0.5$) and is compared with the nearest neighbor (NN) and bilinear (BL) interpolation. Results are expressed in terms of overall average PSNR among the complete test set, at 4 and 8× zoom factor (Fig. 3.4 and 3.5 respectively).

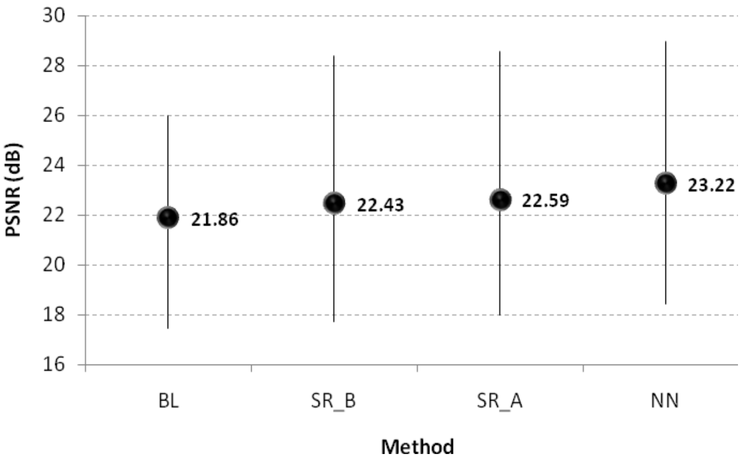


Fig. 3.4. Average PSNR results, 4× zoom factor.

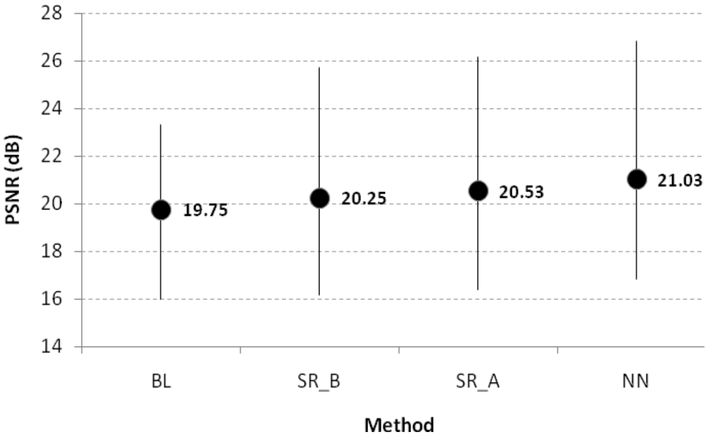


Fig. 3.5. Average PSNR results, 8× zoom factor.

As expected, reconstruction quality decreases as the zoom factor increases. Such behaviour characterizes all methods and derives from the increasing lack of information. In fact, the subsampled frames constitute the only piece of known information for all reconstruction methods.

Objective results show that the proposed method lies between the nearest neighbour and the bilinear interpolation. The superior performance of the nearest neighbour interpolation is easily explained. Since subsampling is carried out through block averaging, nearest neighbour substitution simply assigns the local average to unknown pixels, thus approximating their value with the best estimation in terms of mean square error, thus PSNR.

Then, the proposed method outperforms bilinear interpolation by 0.65dB at 4× zoom factor and 0.89dB at 8× zoom factor on average. It must be observed that the proposed method's performance increases at higher zoom factors if compared to bilinear interpolation.

The results for two sequences are reported in Fig. 3.6 and 3.7.

As previously noticed, nearest neighbor generally outperforms any competing method. Then, PSNR is not fully adequate in providing a reliable quality index. It is a measure that provides an approximate performance indication and cannot be considered as an accurate indication of reconstruction quality. In fact, it does not take into account the issues related to the human visual system and subjective scene interpretation.

Visual results are provided in Figs. 3.8 and 3.9 in order to subjectively evaluate the proposed method. The proposed method results appear visually more pleasant than the competitors. In particular, a detail of the Y component of the "Bus" sequence is reconstructed through NN, SR and BL and is shown in Fig. 3.8 for comparison. The SR reconstruction appears sharper and more defined than the competing techniques.

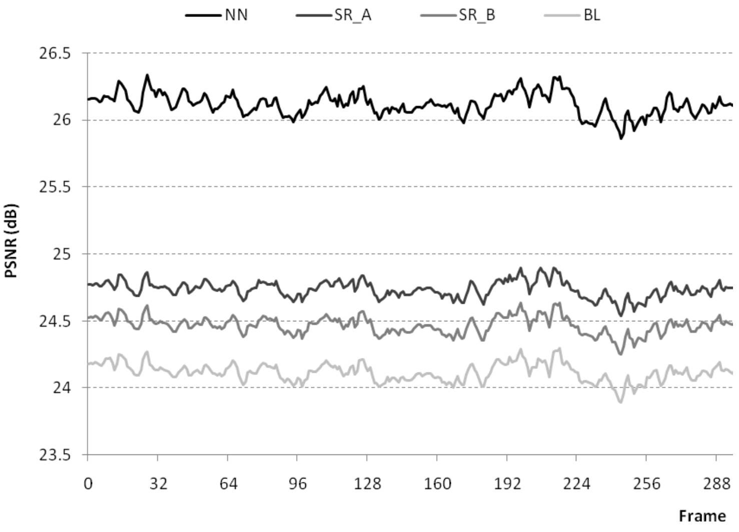


Fig. 3.6. PSNR results for the sequence "Silent" at 4× zoom factor.

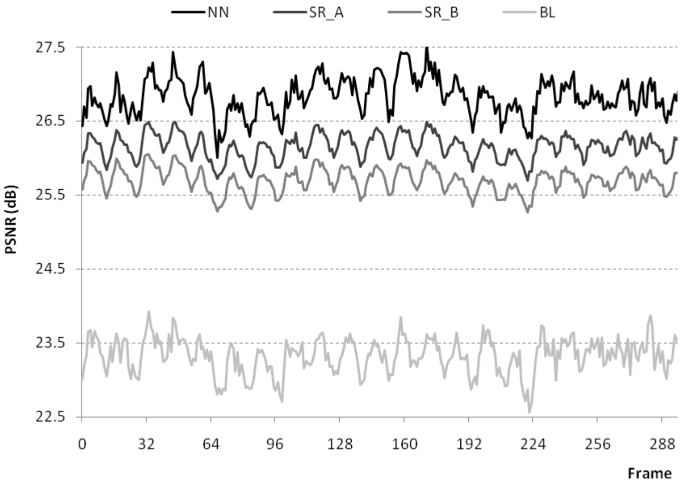


Fig. 3.7. PSNR results for the sequence “Highway” at 8× zoom factor.

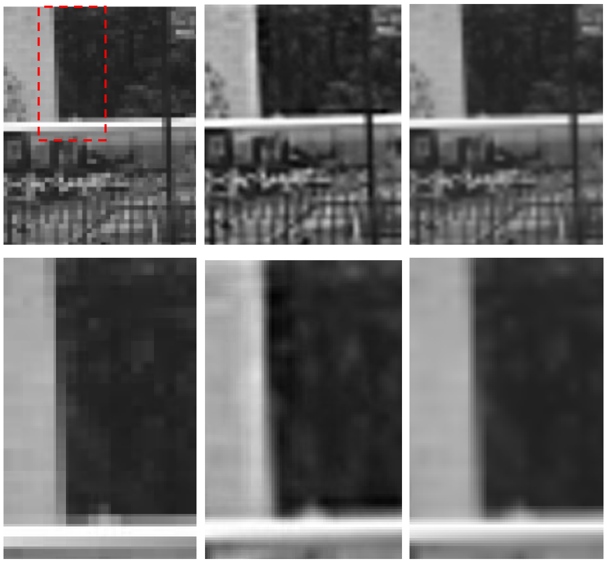


Fig. 3.8. A detail from the sequence “Bus” at 4× zoom factor; from left: NN, SR, BL.

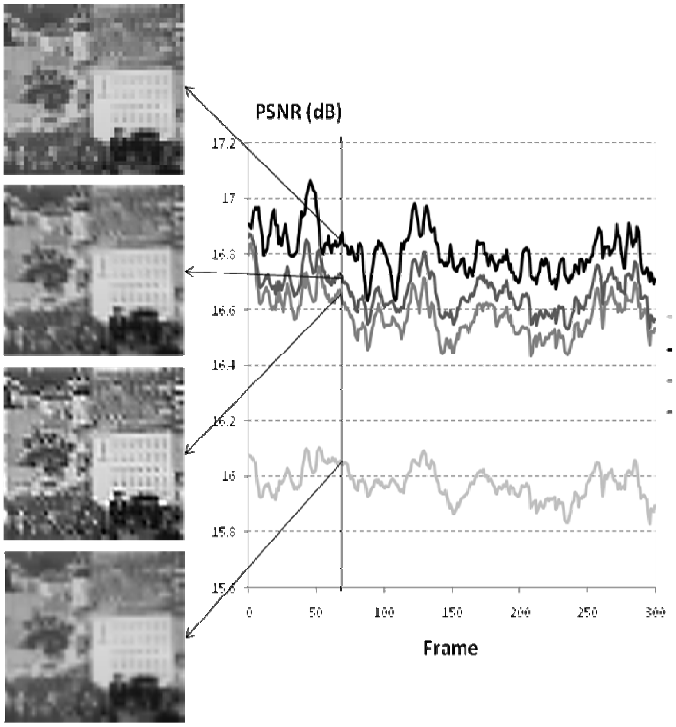


Fig. 3.9. Visual results for the sequence “mobile” at 8× zoom factor.

3.2.6 Conclusions

In this work, an iterative technique for high-resolution reconstruction of low-resolution video sequences has been presented. The proposed algorithm is developed from Peleg and Irani’s works with the generalization of the motion estimation model and modifications to the system architecture. The proposed reconstruction method uses block-based motion estimation and backprojection of the error between the restored frame and the simulated one. Results are promising, especially when considering high zoom factors. Future developments may exploit different motion models and investigate the integration of novel human visual system-based techniques.

3.3 Video Super-Resolution with Adaptive Bilateral Filtering

The current multimedia consumer market is characterized by the advent of cheap but rather high-quality high definition displays, mostly for home theater applications. This trend is only partially supported by the deployment of high-resolution multimedia services, either over the Internet or through satellite channels. To address the resulting disparity between content and display formats, video super-resolution techniques represent a major solution. This subject is addressed in this paper, by exploiting the use of the bilateral filtering. This is a spatial filtering operator that relies on dynamically calculating a FIR kernel which has the major advantages of video content adaptability and edge preserving.

During the last couple of years, the multimedia consumer market has been characterized by the advent of cheap but rather high-quality HD (High Definition) displays, mostly for home theater applications. This process is bound to continue at least in the near future, with the introduction of displays of even higher spatial resolution formats, such as DigitalCinema or UHD/UHDTV (Ultra High Definition). This phenomenon is only partially supported by the deployment of high-resolution (spatial and temporal) multimedia services, either over the Internet or through satellite channels. Indeed, the content generation and distribution sector seems not to be able to keep pace with the display technology, which is characterized by a significant decrease of the cost per pixel. Conversely, the cost of transmitting one bit of video information is not going to decrease, at least when sending it at the quality of service level required by the streaming applications. The advances in the video compression domain, which proceeds by roughly doubling the compression rate every 5 years, do not allow for decreasing such cost significantly. Moreover, older productions need to be either re-mastered or post-processed in order to be broadcasted for HD exploitation. The decoding of low-resolution multimedia content then thwarts the benefits of high-resolution displays and involves the use of appropriate signal processing procedures. Low resolution frames then need to be enlarged through super-resolution techniques, with zooming factors that may increase considerably during the next few years.

The present paper focuses on this problem by proposing a solution which resorts on the use of the bilateral filtering [51]. This is a spatial filtering operator that relies on dynamically calculating a FIR kernel. Edge preserving nature and adaptability are the main advantages of this kind of filter. Whereas it has already been adopted to address the super-resolution problem, its application has been mostly restricted to the case of still-pictures. Herein, we propose its use to tackle the video sequences super-resolution problem and, accordingly, we propose several changes in its use. The first change is related to its extension to the time domain through the use of a group of frames when estimating the super-resolution version of each frame. This operation goes in the direction of both strengthening the local visual information sketch and compensating (thus reducing) the local noise in the current frame with that of previous ones. It may be argued that using frames other than the one to be processed may introduce some distortions due to differences between adjacent frames. However, given an adequately small time window, these differences do not modify significantly the local visual structure, as shown by the high correlation between adjacent frames in Fig. 3.10. This graph plots the average interframe correlation of 10 CIF (Common Intermediate Format, 352×288 pixels, 29fps) test sequences with no less than 300 frames, computed for a window of 31 frames. The correlation curve shows that on average 3 consecutive frames have a correlation higher than 0.9 and 5 consecutive frames are correlated as much as 0.85.

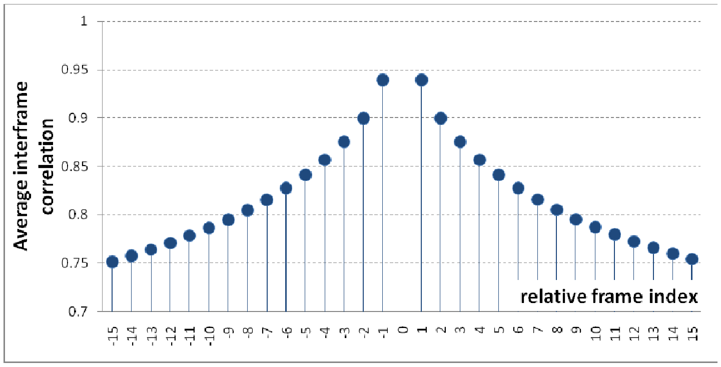


Fig. 3.10. Average interframe correlation.

Instead of relying on a classical motion compensation algorithms, the proposed method implements a 3D sample estimation and filtering.

A second major change we propose is related to the preliminary estimation of the pixel that are added to increase the resolution. While the procedure itself is aimed at estimating these values, these are needed to bootstrap the bilateral interpolation when computing the filter kernels. To address this problem we make use of a gradient based edge-preserving interpolation.

The proposed technique is aimed at reconstructing each HR frame from a limited number of frames extracted from a LR sequence, without any preliminary knowledge of the high-definition data. For any given frame, a sliding time-window determines the set of LR frames (from 2 to N) to be processed in order to produce the output stream. The window is shifted forward to produce successive HR frames of the output sequence, as shown in Fig. 3.11.

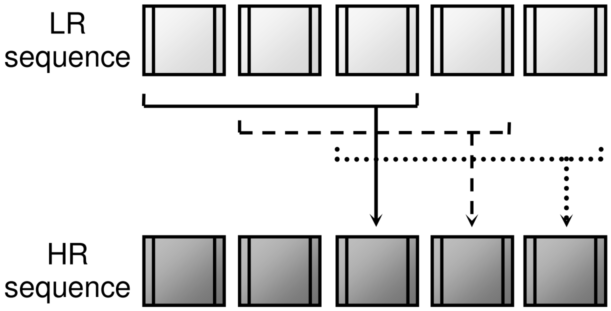


Fig. 3.11. Sliding time-window.

Not to delay the display of the frames, each HR is generated by considering only previous frames. A space-time 3D filter is then applied to such partitioning of the original signal; the filter is developed from the bilateral filter solution with the introduction of sample estimation through local analysis, involving smooth and edge area classification and exploitation.

3.3.1 Background Theory

The proposed interpolation is based on bilateral filtering [51], which relies on dynamically calculating a FIR kernel from known pixels through spatial distance (w_S) and amplitude distance (w_R) weighting contributions:

$$\left\{ \begin{aligned} W_S^{i,j}(h,k) &= \exp\left(-\frac{d^2([i,j],[i-h,j-k])}{2\sigma_S^2} \right) \\ W_R^{i,j}(h,k) &= \exp\left(-\frac{(I(i,j)-I(i-h,j-k))^2}{2\sigma_R^2} \right) \end{aligned} \right\} \qquad (1)$$

where: (i,j) denotes the kernel center; $I(x,y)$ is the signal amplitude at coordinates $p_{x,y}$; $d(x,y)$ is the Euclidean distance function; σ_S^2 and σ_R^2 are the spatial and the amplitude variance, respectively. The kernel coefficients are then computed as follows:

$$W^{i,j}(h,k) = \frac{W_S^{i,j}(h,k) \cdot W_R^{i,j}(h,k)}{\sum_{x,y \in K} W_S^{i,j}(x,y) \cdot W_R^{i,j}(x,y)} \qquad (2)$$

where K represent the set of pixels belonging to the filtering kernel. Fig. 3 illustrates the two weighting contributions and the final kernel shape W. It can be seen that the W_S contribution has a symmetric shape depending only on the distance from the kernel center, while the W_R contribution is modeled by the amplitude distance from the central sample.

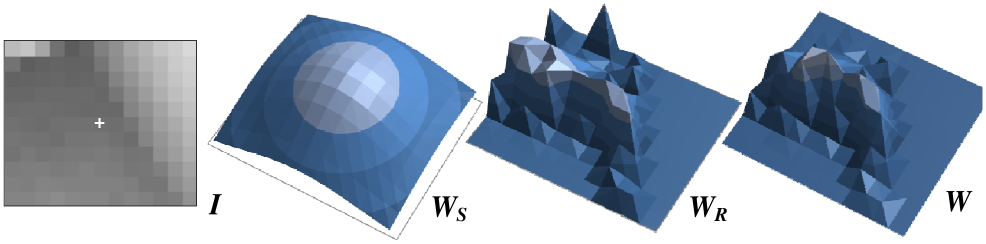


Fig. 3.12 Filter kernel shape related to an edge area.

3.3.2 Super resolution with tridimensional bilateral filter

In the proposed technique, we extend the bidimensional bilateral filter described in the previous section into a tridimensional filter adding the temporal axis. Additionally, we make use of kernel with three equal edges. Given the size of the sliding time-window, N, the linear size of the LR kernel, s_W, and the linear zoom factor, zf, the cubic filter kernel will entail a local lattice with size:

$$N \cdot s_W^2 \cdot zf^2 \tag{3}$$

It can be observed (Fig. 3.13) that only $N \cdot s_W^2$ samples are known from the original signal. The bilateral interpolation then consists in reconstructing the current (unknown) sample through the bilateral formulation.

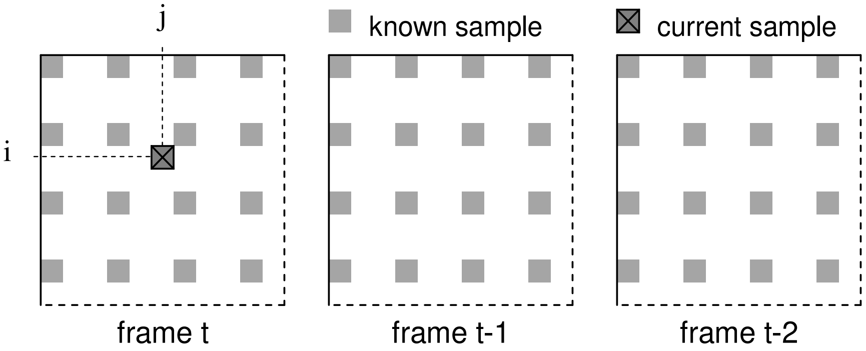


Fig. 3.13. HR image lattice for the kernel support.

However, while the spatial term, W_S, can be easily computed by considering the spatial distances in the HR lattice, the amplitude term, W_R, lacks the definition of the sample value itself. In order to process the signal, such value must be estimated. Given $\hat{I}$, the amplitude estimate, (i, j, t) spatial (intra-frame) and temporal (inter-frame) dimensions respectively, (1) becomes:

$$\left\{ \begin{array}{l} W_S^{i,j,t}(h,k,l) = \exp\left(-\dfrac{d^2([i, j, t], [i-h, j-k, t-l])}{2\sigma_S^2} \right) \\ \hat{W}_R^{i,j,t}(h,k,l) = \exp\left(-\dfrac{\left(\hat{I}(i, j, t) - I(i-h, j-k, t-l)\right)^2}{2\sigma_R^2} \right) \end{array} \right\} \tag{4}$$

In order to estimate the current sample value, a local analysis is performed, based on the LR edge map. The process is graphically described in Fig. 3.14. Both edge magnitude and orientation are firstly computed through a gradient operator. Only strong edges are considered by applying a threshold to the edge magnitude values. For each neighborhood, a linear edge model is derived through the computation of the local edge center of mass and the average edge normal angle:

$$\left\{ \begin{array}{l} i_c, j_c = \dfrac{1}{N_{p_{h,k}}} \sum\sum h_{p_{h,k}}, \dfrac{1}{N_{p_{h,k}}} \sum\sum k_{p_{h,k}} \\ \theta_c = \dfrac{1}{N_{p_{h,k}}} \sum\sum \theta_{p_{h,k}} \end{array} \right\} \qquad (5)$$

with i_c, j_c coordinates of the edge center of mass, θ_c average edge angle, $h_{p_{h,k}}, k_{p_{h,k}}$ coordinates of edge pixels, $\theta_{p_{h,k}}$ edge pixel angle and $N_{p_{h,k}}$ number of edge pixels.

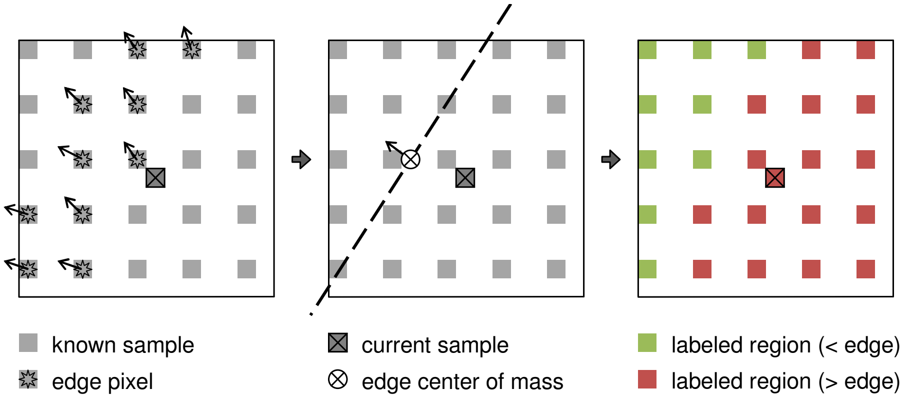


Fig. 3.14. Local neighborhood analysis.

Known samples are then classified as belonging to either the same (SS) or the other side of the edge line (OS) in comparison with the current sample, according to the following rule:

$$\begin{array}{l} \text{if } (i - i_c) < (\arctan\theta_c + \pi/2)\cdot(j - j_c) \text{ then } i \in \mathrm{SS}; \\ \text{otherwise } i \in \mathrm{OS} \end{array} \qquad (6)$$

Once all known samples are classified, the current sample value is computed as the distance-weighted average among the samples from the same class. Notice that the complete process is applied to a time neighborhood of N frames (Fig. 3.15).

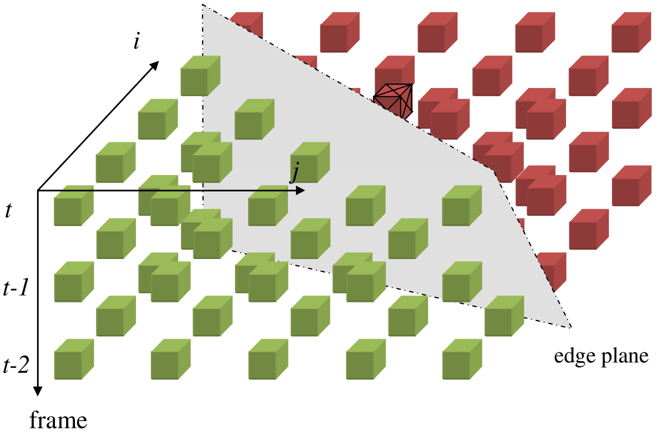


Fig. 3.15. Local neighborhood analysis; space-time structure.

3.3.3 Experimental Results

The proposed method has been evaluated on 12 1280×720 and 1920×1080 4:2:0 YUV video sequences, provided by [59]. The test sequences have been selected with the purpose of presenting a broad range of signal behaviors, in terms of different motion and scene complexity. A subsampled video sequence (Gaussian local filtering) is preliminarily produced from the original video and is used as input sequence for the devised algorithm at any given zoom factor. Test parameters: $N = 3$, $\sigma_S = 10$ and $\sigma_R = 2$. A visual comparison between bicubic interpolation (left) and the proposed method (right) is provided in Fig. 3.16 for three different samples. The proposed algorithm shows a good behavior in both strong and weak edge regions, while highly textured areas are still challenging. Further developments are ongoing in order to deal with such problem through the exploitation of a more precise edge model.

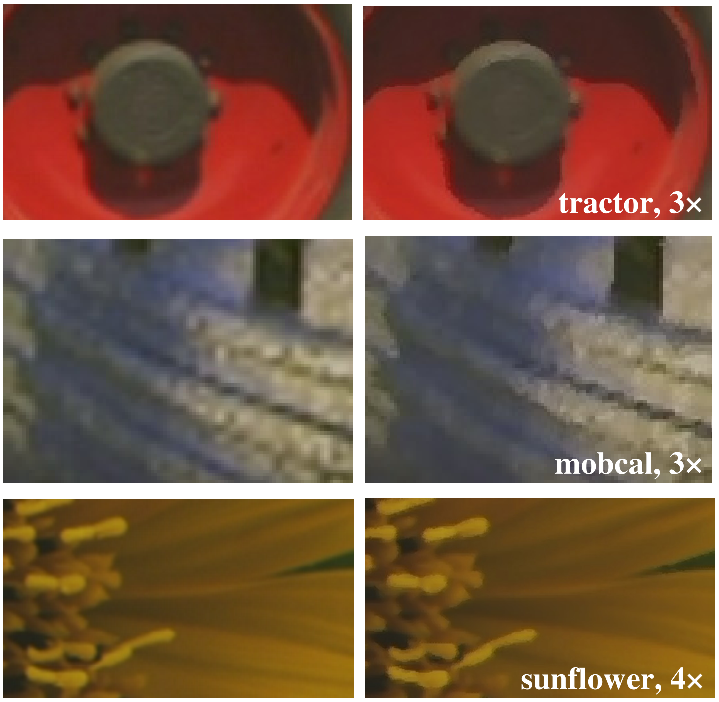


Fig. 3.16. Visual comparison.

3.3.4 Conclusions

A technique for high-resolution reconstruction of low-resolution video sequences has been presented. The proposed algorithm extends the use of the bilateral filter through the exploitation of the space-time domain and the development of edge-based samples estimation, achieving promising results.

Chapter 4 Image Based Positioning System

4.1 Introduction

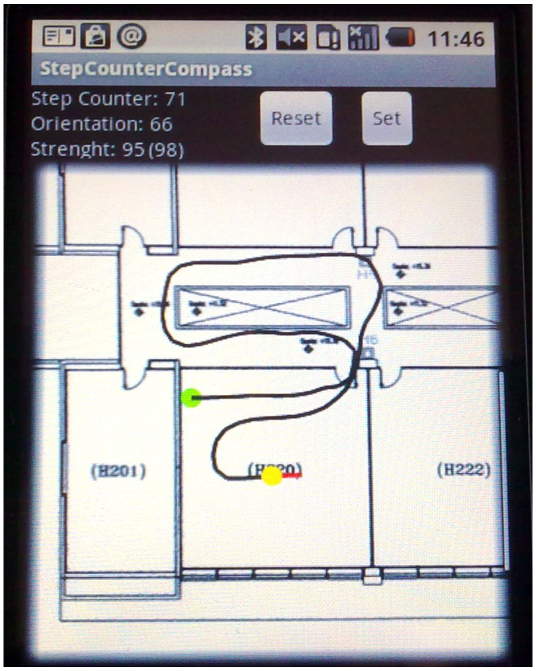
In recent years the mobile telephony market has been characterized by an exponential growth of wearable devices, such as smartphone and PDA in general, equipped with embedded motion (accelerometers) and rotation (gyroscopes) sensors, Internet connection and high-resolution cameras. All the sensing and computing technologies available in a common smartphone makes it ideal for INS (Inertial Navigation System) applications aiming at supporting the navigation of objects and/or users in an indoor environment where common localization systems, such as GPS (Global Positioning System), fail due to severely attenuation or obscuration of the satellite's signal. In fact, the GPS solution is suitable only if at least three satellites are in the line of sight, in the other cases, such as indoor application or urban "canyon", we need to use alternative positioning techniques.

In inertial navigation systems, localization/orientation estimation is source-independent. The user's position is calculated in relation to a known starting position using a dead reckoning algorithm. The whole system makes use of the before mentioned sensors: accelerometers are used to calculate the distance travelled and the gyroscopes/magnetic compass to determine the direction. The uncertainty in the estimated position grows with time from the initial known starting position since the errors introduced by estimating the user/object movements are additive, increasing the total inaccuracy. This demands for a periodic recalibration of the system to reduce the cumulative error. In the following two different approaches are presented. The first one is based on plane homography and affine transformation while the second one is based on SURF. In the first case the considered scenario includes the presence of geo-referenced 2D-tags placed in some known, key positions of the site to be visited. By taking a photo of the tags, the system is able to initialize and subsequently re-calibrate the location data. To improve the calibration accuracy, the focus has been put on computing the exact position of the user (based on the known position of the tag) in terms of orientation and distance from the reference point using plane homography and affine transformation. This allows to correct perspective and projective distortion from the taken photo and derive information about the viewing angle (the user's orientation) and distance between camera and object. In the second work the smartphone's videocamera is used to identify known keypoints, named anchors previously identified and geo-referenced in the building map. For a periodic position fix, an image-based localization system making use of the built-in camera is employed. By developing local feature detection, description and matching between a query image, acquired by the user, and a database containing a collection of geo-referenced images related to the chosen environment, the user's position can be accurately fixed. The proposed solution is based on the SURF (Speed-up robust features), which allows for a quick and effective detection of image features without being affected by the user's viewpoint.

4.2 Inertial Navigation System's Architecture

The considered system is solely based on the capabilities of a common modern smartphone. The data read from the phone's sensors, combined with a reference map of the place and a known starting point, gives the actual position of the user. Hence, there is no need to connect to any external or pre-installed positioning system such as GPS, RFID, or to use WiFi trilateration; solely the dead reckoning technique is used instead. Dead reckoning is the process of estimating the current position of an user based upon a previously known position, upgrading this position upon measured or estimated speeds over elapsed time and course.

The prototype of the proposed system uses the data from the motion sensors embedded in the smartphone to compute the correct position of the user based on a known initial location. The smartphone application is presented in Figure 4.1.


StepCounterCompass

Step Counter: 71
Orientation: 66
Strenght: 95 (98)

Figure 4.1. Screen of the application with a pedestrian route example.

The initial position of the user, the only certain information on which the system relies on for further calculation, is retrieved using the integrated photo camera of the smartphone scanning and decoding a datamatrix (2D barcode) placed aside the map of the floor (see figure 4.2)[60].

Based on the URL encoded in the datamatrix, the application downloads from a dedicated server the indoor vector map for the specific floor together with the initial position of the user on the map (corresponding to the point where the user stands when scanning the datamatrix).

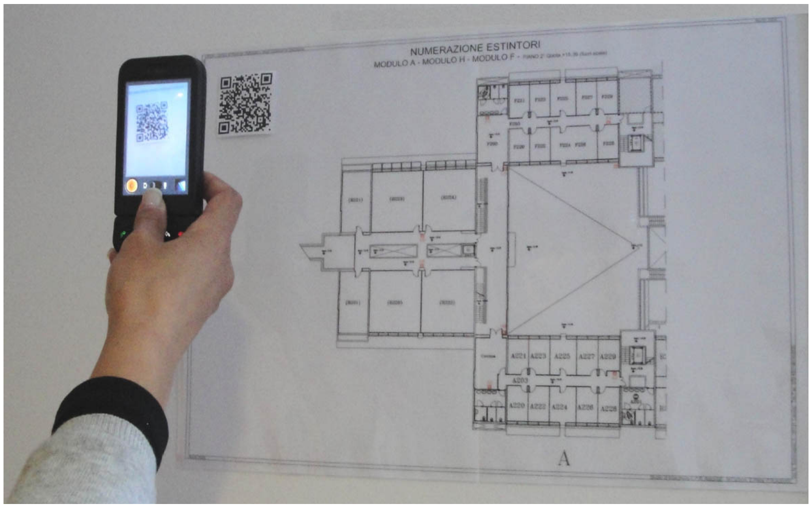


Figure 4.2. User reads a 2D datamatrix to download the map and his starting position.

When the user starts walking, the application draws step by step the position of the user, as a continuous line, over the downloaded map of the building floor.

The application tracks the number of steps taken by the user based on the linear numerical values returned by the smartphone's accelerometers. The acceleration value is the module of the accelerations registered in the x, y and z-axes. One step is detected when this module is above a high threshold (Th_high) and successively is below a Th_low value. The following figure is an example graph of the accelerometer's measurement for a step length of 70cm, having the absolute values $Th_{high} = 109$ and $Th_{low} = 97$.

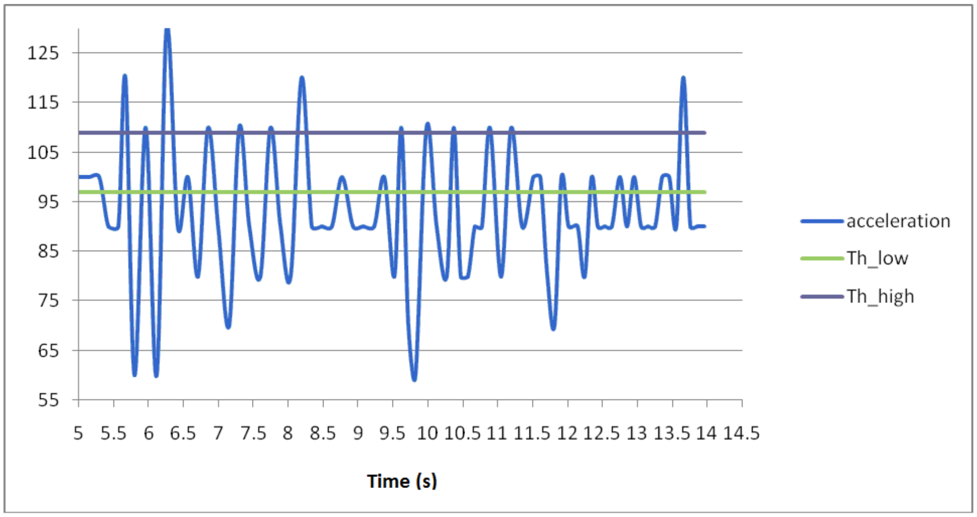


Figure 4.3. Normalized Acceleration

The current orientation of the user is measured by the smartphone’s digital compass (the parameter 'Orientation' in Figure 4.1). The initial orientation, in this implementation, is set when the user scans the 2D barcode, being perpendicular (within a certain angle) to the floor plan. The relative position of the device with respect to the user (e.g. in a pocket) does not influence the dead reckoning estimation. If the device is held in front of the user, the magnetic compass provides the step-by-step heading improving the overall accuracy of the positioning method. This approach does not consider the user’s distance from the object, therefore it’s not too accurate. As showed by first experimental results, before starting the application, the compass needs an accurate recalibration. This recalibration is necessary because the compass is subject to several errors: initially it has an inaccuracy of maximum 5 degrees, depending also on the used device and on the presence of electromagnetic interferences.

The step counter module based on the accelerometer data was tested and validated after thorough tests, performed in an indoor environment using both men and women with different physical features. The mean placement error was 3,8% on a series of 20 runs consisting of an average step count of 40 steps. In order to better estimate the user position the investigated technique based on plane homography is presented in the following section.

4.2.1 Calibration of Inertial Navigation System

In order to reduce the cumulative error an early development of an Indoor Navigation System based solely on the capabilities of a typical modern smartphone equipped with accelerometers, compass, camera and Internet connectivity has been proposed. The user initially takes a photo of a geo-referenced 2D-bar code in order to acquire the map of the building and the initial position. The system then estimates the movement calculating the number of user’s steps from the starting point using the accelerometers and the direction using the compass. The considered scenario includes the presence of geo-referenced 2D-tags placed in some known, key positions of the site to be visited. By taking a photograph of the tags, the system is able to initialize and subsequently re-calibrate the location data. To improve the calibration accuracy, we have focused on the problem of computing the exact position of the user (based on the known position of the tag) in terms of orientation and distance from the reference point using plane homography and affine transformation [61]. This allows us to correct perspective and projective distortion from the taken photo and derive information about the viewing angle (the user's orientation) and distance between camera and object.

Inertial Navigation System are used in many different kind of application involving moving objects, including vehicles, such as for example aircraft and submarines for navigation purpose. Recently, some research has proposed its use as assistive mobility technology for people with some kind of disabilities. Furthermore, indoor navigation can support commercial activities such as the retrieval of products in a large mall, but can also be deployed for security reasons: evacuation of complex buildings, route identification for visitors etc.

Based on the first experimental results presented in the previous paragraph, we identified the need to correct the data returned by the smartphone's sensors, and also to point more precisely the user's initial position.

Therefore, in this section we present the technique that we are investigating for improving the calibration of the system while acquiring the image of the 2D tags mentioned before. The problem we are focusing on is determining the user's position relative to the 2D tags in terms of the precise distance and orientation angle.

We are making use of plane homography techniques: in a central projection camera model, a three-dimensional point in space is projected onto the image plane by means of straight visual rays from the point in space to the optical centre. Mathematically this process can be described using a $3\text{x}4$ projection matrix P, which takes a point in 3-D space in homogeneous coordinates $(X,Y,Z,1)^T$ and transforms it into a point on the 2-D image plane $(x,y,1)^T$.

$$\lambda \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = [P] \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \qquad (1)$$

The projection matrix P can be computed from the internal and external camera parameters:

$$P = K\left[R|T\right] \qquad (2)$$

where K is a 3x3 upper triangular matrix, called the camera calibration matrix, including the intrinsic camera parameters (focal length, aspect ratio and skew) and $[R\,|\,T]$ defines the Euclidean transformation between camera and world coordinates (in general rotations followed by translations), including the external camera parameters, i.e. its position and orientation. In the case where planar surfaces are imaged ($Z = 0$), the transformation is called a plane-to-plane homography:

$$\lambda \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = [H] \begin{pmatrix} X \\ Y \\ Z = 0 \\ 1 \end{pmatrix} \qquad (3)$$

The 3x3 transformation matrix, usually called the homography matrix H, has a simpler form than P, but it can be also reduce to:

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} = [K]\left[r_1 \quad r_2 \quad t\right] \qquad (4)$$

where K is the camera's matrix, r_1 and r_2 are the correspondent columns of the rotation matrix R and $t = -RC$ with C the camera center.

For this particular case we are dealing with the acquisition of a planar surface. Figure 4.4 shows the mapping between a 2-D point x' in the object plane π' where the tag lies into a 2-D point x in the image plane π that represents the image acquired by the camera.

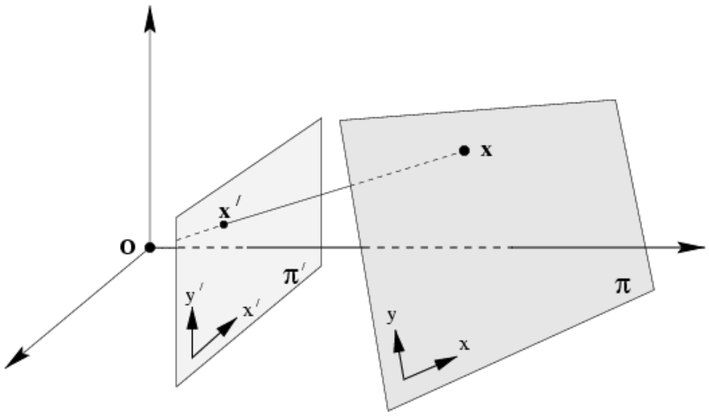


Figure 4.4. Mapping between planes

This process can be described mathematically by a homography matrix H :

$$P_i' = HP_i \qquad (5)$$

where P and P' are $3x1$ vectors that could correspond to the images of the same points, the former in the plane of the tag and the latter in the plane of the image, while H is the transformation matrix.

If the homography between a plane in the scene and the plane of the image is known, then the image of the planar surface can be rectified into a front-on view. Given four points on the scene plane, with no more than any 2 points collinear, and their corresponding positions in the image (8 equations), H is uniquely determined. Let $P_1'(x'_1, y'_1)$, $P_2'(x'_2, y'_2)$, $P_3'(x'_3, y'_3)$ and $P_4'(x'_4, y'_4)$ be the four corner points of the rectangular object and $P_1(x_1, y_1)$, $P_2(x_2, y_2)$, $P_3(x_3, y_3)$ and $P_4(x_4, y_4)$ their projections obtained using a plane homography transformation.

Corresponding points in two images related by homography are then:

$$x_i' = \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + h_{33}}$$

$$(6)$$

$$y_i' = \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + h_{33}}$$

Our objective is to remove affine and projective components in order to obtain a similarity transformed such for example a rotated, scaled and/or translated version of the original image.

To do this, we firstly have to identify the four corners in the taken photo, then, using plane homography transformation, map each vertex of the quadrilateral to the corresponding vertex in the known rectangle. Using equations (6) we can find the coefficients of the homography matrix H and finally rectify the image to the frontal view. Once the frontal view is recovered from the knowledge of the calculated homography's matrix coefficients, we can decompose the H matrix, using QR Factorization, in its orthogonal ($\left[r_1 \quad r_2 \quad t\right]$) and upper triangular matrix $\left[K\right]$.

From the knowledge of the orthogonal matrix we can determine the tilt angle φ (rotation around the x-axis), the roll angle ψ (rotation around the y-axis), the pan angle θ (rotation around the z-axis) and the translation along the three axes, thus the orientation and position of user/camera in the scene. From $\left[K\right]$, given the focal length of the camera embedded in the smartphone, and given the real dimensions of the 2D datamatrix, we can calculate the distance from the datamatrix.

Using the results of the calculus, the initial position of the user and his orientation relative to the scanned 2D datamatrix can be more precisely computed, as being of crucial importance for the correct evaluation of the data subsequently generated by the smartphone's sensors, especially in terms of orientation.

4.2.2 Conclusions

The presented solution for a pedestrian indoor navigation system has been developed on a modern Smartphone and was tested in a real indoor environment, measuring the encountered errors. The application of the plane homography technique to the indoor navigation problem has been investigated in order to derive additional information about the relative orientation and distance of the user to the reference point. Based on this supplementary data we are trying to reduce the inherent errors of the dead-reckoning technique.

4.3 An Image based positioning system using SURF

Among the wide number of indoor navigation solutions, we propose a system capable to localize a user on the basis of the capabilities of a modern smartphone equipped with camera, digital compass, accelerometer and WiFi connection. The only external infrastructure is given by some 2-dimensional barcodes positioned in key points.

In a typical scenario a user needs to move from place A to place B in an unknown indoor environment. The initial position of the user is retrieved by scanning and decoding a geo-referenced datamatrix (2D barcode) placed aside the map of the floor with the embedded phone's camera. The maps with the barcode are assumed to be hanged on the wall at the interest points. Based on the URL encoded in the datamatrix, the application downloads from a dedicated server the digital indoor vector map for the specific floor together with the initial position of the user on the map (corresponding to the point where the user stands when scanning the datamatrix). The user's initial position is more precisely defined in term of distance and orientation angle from the reference QR code using plane homographic techniques. When the user starts walking, the application draws step by step the position of the user, as a continuous line, over the downloaded map of the building floor. The application tracks the number of steps taken by the user based on the numerical values returned by the smartphone's accelerometers as described in Section 4.2.1.

The heading is retrieved by considering the output of the magnetometer. Taking in consideration that the magnetometer retrieves the magnetic north with respect to the phone's current orientation which might be diverse form the walking direction of the user, the need for a compensation of the heading arises. This compensation is performed by analyzing the position of the phone with respect to the user starting from an initial known position.

On the basis of the corrected heading and the number of steps taken, the application deduces that user is near to some anchor points and suggests him to recalibrate the system in order to reduce the position error. Thus the user takes a photo of the closest anchor point, sends it to the server and waits for the response that will show the most probable position in the map building on the phone's display.

By developing local feature detection, description and matching between a query image, acquired by the user, and a database containing a collection of geo-referenced images related to the chosen environment, the user's position can be accurately fixed.

The proposed solution is based on the SURF (Speed-up robust features), which allows for a quick and effective detection of image features without being affected by the user's viewpoint. The INS system can be recalibrated (position fix) by taking photo of anchor points (nodes with a known position) present in the indoor environment.

4.3.1 Background Theory

4.3.1.1 Acceleration Sensors

The integrated accelerometers in modern smartphones are tri-axial devices which allow the detection of accelerations forces in m/s^2 along the X, Y and Z axes. The values of the acceleration are positive or negative based on the direction in which the phone is moving and based on the position of the phone. When the device is lying still in a flat position, the accelerations on the three axes are:

$$\begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -g \end{bmatrix}, \qquad (7)$$

where g is the earth's gravitation. In fact, using the presence of gravity distributed on the three axes, the orientation of the device can be calculated using the modulus of the accelerations |am| given by

$$|a_m| = \sqrt{a_x{}^2 + a_y{}^2 + a_z{}^2}, \qquad (8)$$

and the following equations:

$$\alpha = \cos^{-1}\left(\left|\frac{a_x}{a_m}\right|\right), \beta = \cos^{-1}\left(\left|\frac{a_y}{a_m}\right|\right) \qquad (9)$$

The angles α and β are the angles which the device is forming with the X and Y axes. The scaled values of α and β give the angles for roll and pitch, as illustrated in Fig 4.5. For the smartphone lying still along the Y axis on a flat horizontal surface, the pitch and roll are equal to zero, changing when moving from 0° to 360°.

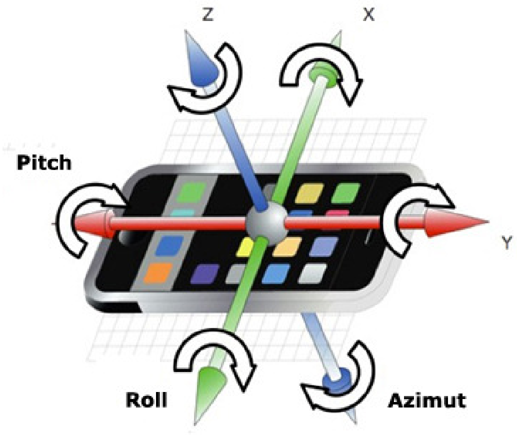


Fig. 4.5. Axes and rotation angles of a smartphone

4.3.1.2 Magnetic Sensor

The azimuth γ from figure 1 represents the angle between the magnetic north and the Y axis of the smartphone with the display heading up. Just like a digital compass, the values for the azimuth are between 0° and 359°, with 0° for the magnetic north, 90° for east and so on. The azimuth value returned by the magnetic sensor of a smartphone is highly susceptible to electromagnetic interference and is also quite instable for typical devices, emerging the need for a periodic recalibration, which can be performed by rotating the smartphone in a 8-like pattern [62].

4.3.1.3 SURF

SURF is an efficient scale and rotation invariant interest point detector and descriptor. It allows for quick and effective feature detection even against different image transformations like image rotation, scale illumination and small viewpoint changes.

Much of the performance increase can be attributed to the use of an intermediate image representation, known as the Integral Image that can be rapidly computed from an input image [50]. This section shows a brief summary of its construction process.

a. Interest point detection

SURF is a Hessian matrix based interest point detector. It searches for blob-like structure at locations where the determinant of this matrix is maximal. Given a point $X = (x, y)$ in an image $I(x, y)$, the Hessian matrix $H = (X, \sigma)$, as function of both space X and scale σ, is defined as follows:

$$H(X, \sigma) = \begin{bmatrix} L_{xx}(X, \sigma) & L_{xy}(X, \sigma) \\ L_{xy}(X, \sigma) & L_{yy}(X, \sigma) \end{bmatrix}, \qquad (4)$$

where $L_{xx}(X, \sigma)$ refers to the convolution of the second order Gaussian derivative $\frac{\partial^2 g(\sigma)}{\partial x^2}$ with the image at point $X = (x, y)$ and similarly for $L_{yy}(X, \sigma)$ and $L_{xy}(X, \sigma)$. These derivatives are known as Laplacian of Gaussians. The approximated determinant of the Hessian represents the blob responses at location $X = (x, y)$ in the image. In order to detect interest points over different scale a non maxima suppression in a 3 x 3 x 3 neighbourhood is applied. To do this each pixel in the scale-space is compared to its 26 neighbours, comprised of the 8 points in the native scale and the 9 in each of the scales above and below. Finally the maxima of the determinant of the Hessian matrix are then interpolated in both space and scale to sub-pixel accuracy.

b. Interest point descriptor

The SURF descriptor describes the distribution of pixel intensities within a scale dependent neighbourhood of each interest point detected by the Fast-Hessian. Integral images in conjunction with Haar wavelets are used in order to increase robustness and decrease computation time. Haar wavelets are used to find gradients in the x and y directions. The first step in descriptor's extraction consists of fixing a reproducible orientation based on information from a circular region around the interest point. Then, a scale dependent window aligned to the selected orientation is constructed and a 64-dimensional vector (SURF descriptor) is extracted from it. The dominant orientation is estimated by calculating the sum of all responses within a circle segment covering an angle of $\pi / 3$ around the origin. At each position, the two summed x and y responses are used to form a new vector.

The longest vector defines the orientation of the interest point. The first step for the extraction of the descriptor is to construct a square region aligned with the selected orientation around the interest point. It contains the pixels which will form entries in the descriptor vector and is of size20σ, where σ refers to the detected scale. A further division into $4x4$ regular sub regions is performed within each Haar wavelets of size 2σ, calculated for $5x5$ regularly spaced sample points. Hence, each sub-region has a four dimensional descriptor vector, thus concatenating this for all 4$x4$ sub-regions a descriptor vector of length 64, invariant to different image transformation is obtained.

c. Descriptor Matching

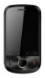
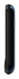
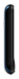
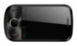
The descriptor matching is performed by implementing the so called One to One algorithm [63]. Given two sets of descriptors $\{P\}$ and $\{Q\}$ extracted from a pair of images $(I_{1,}I_2)$, it returns pairs of closest descriptors using an Euclidean metric $\rho(P, Q)$.

4.3.2 Heading correction

While the step counter presented in Section 4.2.1. and based on the modulus of the output of the tri axial accelerometer produces satisfactory results, problems arise for accurately determining the heading of the user. These problems arise due to the fact that the output of the magnetic sensor is related to a smartphone in a flat position, heading in the same direction as the user. If the smartphone gets in a different position, for example used for talking on the phone or placed in a pocket, the change in position will be erroneously intended as a heading change.

To compensate the heading of the smartphone for position changes relative to the user, we developed a position classifier based on the interpretation of the pitch, roll and relative azimuth values as defined in the third section. For simplifying the classifier, we assumed only 90° rotations for each of the three axes, resulting in 8 possible positions for each axe (4 for the positive values and 4 for the negative values), for a total of 24 positions. The classification is based on the values of pitch and roll, with a tolerance of +/- 30°. An excerpt of the 24 positions is presented in table 1, for a smartphone held with the screen vertically in front of a standing user, vertically on the left and right side and laterally rotated.

TABLE I
EXCEPT FOR THE SMARTPHONE POSITIONS BASED ON THE PHONE'S ANGLES WITH THE THREE AXES

Pos.	***a***	***β***	**γ**
	$60^\circ < a < 120^\circ$	$330^\circ < \beta < 30^\circ$	$\gamma_{ref}+330^\circ < \gamma < \gamma_{ref}+30^\circ$
	$150^\circ < a < 210^\circ$	$330^\circ < \beta < 30^\circ$	$\gamma_{ref}+120^\circ < \gamma < \gamma_{ref}+60^\circ$
	$330^\circ < a < 30^\circ$	$330^\circ < \beta < 30^\circ$	$\gamma_{ref}-120^\circ < \gamma < \gamma_{ref}-60^\circ$
	$150^\circ < a < 210^\circ$	$60^\circ < \beta < 120^\circ$	$\gamma_{ref}+120^\circ < \gamma < \gamma_{ref}+60^\circ$

The reference azimuth γref is recorded in the moment when the user is initially scanning the geo-referenced datamatrix and is presumed known, taking into consideration that the user has to stand in front of the data matrix to perform the scan.

Starting from this point, the heading of the user is calculated based on the actual azimuth values given by the smartphone magnetometer corrected by values corresponding to the calculated position of the smartphone.

4.3.3 Position Fix Using Anchor Points

The periodic position fix is addressed by developing a local feature detection, a description and a matching algorithm between a query image (acquired in real time by the user) and a database containing a collection of geo-localized images. The entire process can be basically divided in offline and online phases.

The offline phase, specific to each building, has to be executed only once (or when new anchor points need to be introduced in the indoor environment), resulting in the creation of a database. The data acquisition block can be seen as a sort of calibration: a certain amount of anchor points/locations will be chosen, depending on the size and layout of the building. At each of these locations, a subset of n photos from fix distance and different direction is taken in order to maximize the probability to have a true match. Once collected, every image is process with the SURF algorithm to extract significant features, which are then coded in a descriptor vector. The created database represents a collection of anchor points at different locations in the building, taken under various illumination condition (light on/off) and from different viewpoints (frontal or lateral view). In particular we choose as anchor point internal/external door and gate, lighting system and air conditioning system.

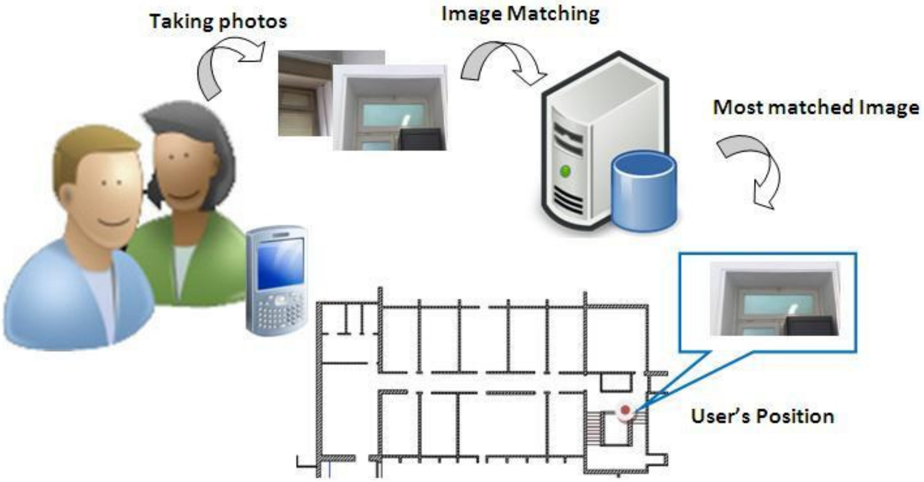


Fig. 4.6. Estimating position using anchor points

During the online phase, as shown in Fig. 4.6, a user who wants to know his current position collects an image of his surrounding on the basis of the anchor points proposed by the system and sends the captured query image to the database for localization upon the map. To perform positioning, the algorithm investigates how similar the query image is to each image in the database by extracting and comparing its features with those of other images in the database. Finally, the application returns the image of the most probable location fixes the user position on the display.

4.3.4 Experimental Results

The proposed algorithm was tested on an iPhone 3 GS with a 600 MHz ARM cortex Processor and a 3Mp built-in camera. The feature extraction functionalities have been implemented making use of the the OpenCV library. All tests were run on an Apple Mac Book Pro Intel Core 2 Duo machine with 2.4 GHz, 4 GB.
The tests for evaluating the feasibility of the position model and the corrections on the heading were performed using a reference circular path. This known path was run ten times holding the smartphone with its Y axis straight towards the walking direction .

For a number of 16 points on this path, the azimuth was calculated as the mean azimuth from the 10 runs.

Subsequently, the known path was covered in a similar manner but moving the smartphone with respect to the user's body in a series of 5 previously known typical positions: front chest pocket, side pocket, rear trouser pocket, left ear, right ear. The changes in position of the smartphone were performed at known instants of time, in order to be able to synchronize the values with the test runs. The known path was covered performing the same position changes for 10 consecutive times.

Table 2 shows the mean azimuth errors in degrees for the 16 known points on the path. The 5 grayed columns represent the points where the position changes took place. The azimuth values in these columns are the corrected ones based on the method presented in the fourth section.

TABLE II

AZIMUTH ERRORS FOR A KNOWN PATH

Point	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Error (°)	5	10	18.2	7.5	20	6.2	5.5	13.4

Point	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Error (°)	7.2	15.6	4.3	8.9	12.2	18.1	8.3	11.4

As it can be noticed the azimuth errors for the points with no smartphone position change are less than 10°, while the errors for position changes are higher, but still in an acceptable range, not more than 18°.

We performed the tests in a building of the Campus of the University of Cagliari, where 35 anchor points have been chosen. For each of these anchors we captured 3 photos, at a distance of around 3 meters and with 3 different viewpoints: 0° (frontal view), + 45° and – 45°. Thus we have a database made of a total of 105 images. The number of features could vary a lot among images, from several hundreds to few thousands, as shown in the following test images.

The first test was carried out on an image with 65 features belonging to air conditioning system category as shown in Fig. 4.7.

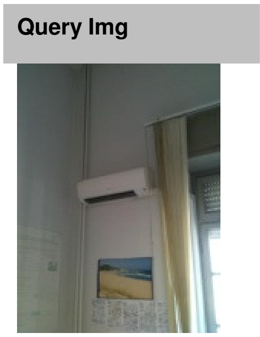
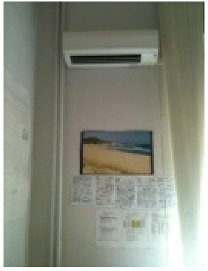
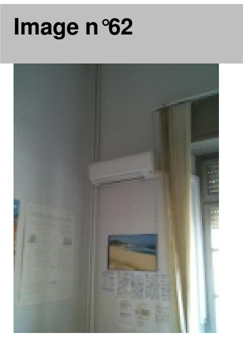
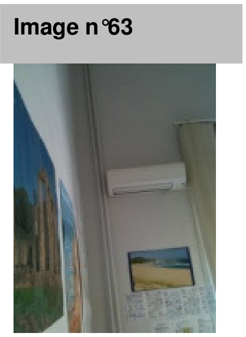
Query Img	Image n°61	Image n°62	Image n°63
			

Fig. 4.7 Query image (left) and selected images from the DB.

The graph in Fig. 4.8 shows the number of features that match between the query image and each photo in the entire database.

The blue line shows the results when no threshold value in the One to One algorithm is applied. The yellow, green and red lines correspond respectively to Th=0.5, Th=0.7 and Th=0.9. As it can be noted, if we consider as query image a generic view of image n° 63 (with 65 features), the highest number of correct matching feature has been obtained for Th=0.5 and Th=0.7. For the mentioned threshold values, we find three maxima in correspondence of images n° 61, n° 62 and n° 63 in the DB. The graph in Fig. 4.8 shows how the algorithm correctly selects the references image and finally identifies the most matched image as the query image (n°63).

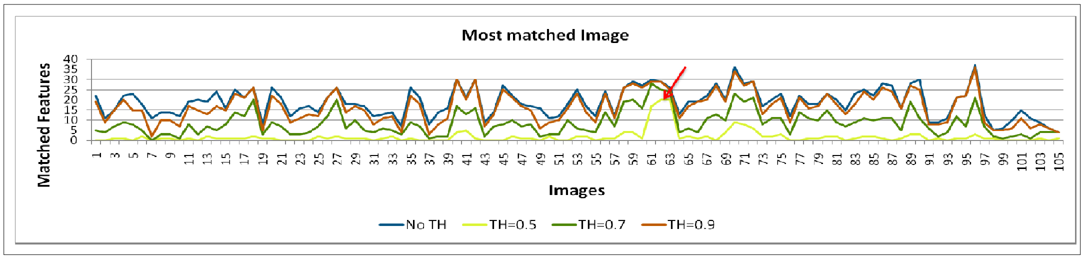


Fig. 4.8 Correct match

4.3.5 Conclusions

In this paper we presented an indoor localization solution that use only the capabilities of a modern PDA equipped with a high resolution built-in camera, internet connection, motion and magnetic sensors, an image recognition system and a map with several geo localized images of the building. The proposed prototype is based on SURF algorithm for feature extraction and description, the One to One algorithm for descriptors matching and on processing of accelerometer and magnetometer data for counting steps and calculating heading. Several tests were carried out and the results are promising. To ameliorate the real time of the entire project, future developments will consist in the integration of a plane homographic technique. This will allow for a better estimation of the user position in terms of view angle and distance from anchor point. In addition, more refined processing techniques for motion and heading data will be employed.

Conclusions

The research developed during these three years has concerned the field of signal processing, with particular attention to image and video analysis. The continuous development in signal processing techniques and the amazing growth in computing power have led to an incredible proliferation of applications based on the use of multimedia data, such as image and video, both for desktop platforms and wearable devices.

In this thesis, two main issues have been addressed: the first one has concerned image and video reconstruction/restoration for multimedia data fruition over the web while the second one was related to image analysis for location-based systems in an indoor scenario.

In the context of multimedia data fruition, many important factors had to be considered, such as bandwidth limitation, availability and easy retrieval of high quality data, service interactivity, etc. In our opinion and to the best of our knowledge, modern signal processing techniques, such as super resolution, can represent the most effective solution, since it allows for restoring the original spatial resolution from low-resolution compressed data. In this scenario, different solutions have been presented both for image and video sequences.

The proposed super resolution techniques were compared with classical interpolation method using objective metrics (PSNR), subjective metrics (SSIM) and providing visual results. As regards of image reconstruction solution, objective results (PSNR) showed that it was comparable with classical interpolation techniques, such as bicubic and bilinear interpolation, whereas it was able to outperform the bilateral filtering interpolation. As regards of video reconstruction solutions, the experimental results have shown good behavior in both strong and weak edge regions, especially when considering high zoom factors, while highly textured areas were still challenging. In all tests carried out we have noticed an apparently unpredicted performance of the nearest neighbour interpolation. It was easily explained: since subsampling was carried out through block averaging, nearest neighbour substitution simply assigned the local average to unknown pixels, thus approximating their value with the best esteem in terms of mean square error, thus PSNR. From the previous considerations, PSNR seemed apparently inadequate in providing a reliable quality index. It is a measure that provides only a rough performance indication and cannot be considered as an accurate indication of reconstruction quality. For these reasons we have also investigated the use of a subjective metrics (SSIM) that take into account the image quality on the basis of the degradation of structural information, providing a better visual quality esteem than PSNR. However, also SSIM cannot still be taken as a perfect indication of the reconstruction quality process. The investigated super resolution techniques, both for image and video sequences, have been proved to be able to meet the requirements of bandwidth limitation, service interactivity etc. Thanks to the reconstruction process we will able to sent over the network only thumbnail version of the images or video frames, thus reduce bandwidth usage and ensure high quality data at the receiver side.

The second topic was related to the implementation of an image based positioning system for an indoor navigator. As it is known, common localization systems, such as GPS (Global Positioning System), fail in indoor environment, thus the need to investigate alternative positioning techniques Typical approaches are commonly based on the use of external infrastructure such as RFID, WiFi etc These solutions offers very high precision but are affected by high costs, thus the need to investigate the use of alternative techniques, such as image based solutions. In this research activity the focus has been put on the use of image analysis techniques for localization purposes, with main attention to image rectification methods and image recognition using the SURF (Speed-up robust features) algorithm.

We have considered a scenario in which a user needs to move from place A to place B in an unknown indoor environment. In the developed system, the initial position of the user is retrieved by scanning and decoding with the embedded phone's camera a geo-referenced datamatrix (2D barcode) placed aside the map of the floor. In the proposed solution, the user's initial position was more precisely defined in term of distance and orientation angle from the reference QR code. Using plane homographic techniques we were able to derive information about the viewing angle (the user's orientation) and distance between camera and object. A series of intensive tests have been carried out in a real indoor environment, taking into account codes of different sizes (both in terms of different amount of encoded data and in terms of physical dimension) acquired from different viewpoints and distance. The tests have shown how the proposed algorithm can correctly identify the viewing angle of the user and the distance from the acquired QrCode in all cases where the inclination is not too accentuated for the proper identification of the code itself.

In a later work, we dealt with the problem of fixing the position of the user by developing a local feature detection, description and matching algorithm between a query image, acquired by the user, and a database containing a collection of geo-referenced images (anchor points with known position) related to the chosen environment. The proposed solution has been evaluated considering a large database of images characterized by a fair number of features, taken from different viewpoints and representing different subject. The tests have shown how the algorithm, once the correct threshold for the matching phase was established, is able to recognize the correct image in most cases. The processing time was proved to be crucial for real time applications. It depends strongly on the database dimension, thus we have to significantly reduce the search during the matching phase in order to avoid query on the entire image's database. To face the problem, each image in the database has been selected in order to be as different as possible in the feature space, although representing a similar subject (e.g., doors in the same building are usually quite similar). and generic elements around the principal subject (e.g., plates, cabinets, air conditioners) in a photo were used to further distinguish them. In addition, images in the database have been catalogued as belonging to predefined categories in order to restrict the search to a smaller number of images.

In conclusion, the research carried out has shown how modern signal processing techniques can be successfully applied in different scenarios, from image and video enhancement up to image recognition for localization purpose, providing low costs solutions and ensuring real-time performance.

REFERENCE

[1] R.Y. Tsai and T.S. Huang, “Multiframe image restoration and registration,” in Advances in Computer Vision and Image Processing, R.Y. Tsai and T.S. Huang, Eds., vol. 1, pp. 317–339. JAI Press Inc., 1984.

[2] M. Irani and S. Peleg, “Super Resolution From Image Sequences,” in Proc. of the 10th Int. Conf. on Pattern Recognition, Atlantic City, NJ, vol. 2, pp. 115–120, June 1990.

[3] H. Greenspan, C. Anderson, S. Akber, “Image enhancement by nonlinear extrapolation in frequency space,” IEEE Trans. on Image Processing, vol. 9, no. 6, 2000.

[4] B. Morse, D. Schwartzwald, “Image magnification using level set reconstruction,” Proc. Computer Vision and Pattern Recognition (CVPR), pp. 333-341, 2001.

[5] R.R. Schultz, R.L. Stevenson. “A Bayesian approach to image expansion for improved definition,” IEEE Trans. on Image Processing, vol. 3, no. 3, pp. 233–242, 1994.

[6] C. Bouman, K. Sauer, “A Generalized Gaussian Image Model for Edge-Preserving MAP Estimation,” IEEE Trans. on Image Processing, vol. 2, no. 3, Jul. 1993.

[7] S.D. Bayarakeri, R.M. Mersereau, “A New Method for Directional Image Interpolation,” Proc. Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP), Detroit, MI, vol. 24, 1995;

[8] K.P. Hong, J.K. Paik, H. Ju Kim, C. Ho Lee, “An Edge-Preserving Image Interpolation System for a Digital Camcorder,” IEEE Trans. on Consumer Electronics, vol.42, no.3, Aug. 1996.

[9] W. T. Freeman, E. C. Pasztor, “Learning to estimate scenes from images,” in M.S. Kearns, S.A. Solla, D.A. Cohn, editors, Adv. Neural Information Processing Systems, vol. 11, Cambridge, MA, 1999. MIT Press.

[10] W.T. Freeman, E.C. Pasztor, O.T. Carmichael, “Learning lowlevel vision,” Intl. J. Computer Vision, vol. 40, no. 1, pp. 25–47, 2000.

[11] M.F. Tappen, B.C. Russell, W.T. Freeman, "Efficient Graphical Models for Processing Images," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA, Jun. 2004.

[12] S. Battiato, G. Gallo, F. Stanco, "A Locally-Adaptive Zooming Algorithm for Digital Images,," Image Vision and Computing Journal, Elsevier Science Inc., vol. 20, no. 11, pp. 805-812, Sep. 2002.

[13] L. Rodrigues, D. Borges, L. Goncalves, "A Locally Adaptive Edgepreserving Algorithm for Image Interpolation," Proc. SIBGRAPI'02, page(s): 300-305, 2002.

[14] Shujun Fu, "Adaptive Image Interpolation Using Coupled Bidirectional Flow," Proc. IEEE Intl. Conf. on Image Processing (ICIP), vol. 2, pp. II - 970-3, Genoa, Italy, Sep. 2005.

[15] R.K. Martin, K.N Plataniotis, "Digital Camera Zooming Based on Unified CFA Image Processing Step," IEEE Trans. on Consumer Electronics, vol. 50, no. 1, Feb. 2004.

[16] T. Yan, L. Bin, L. Tao, "A Local Image Interpolation Method Based On Gradient Analysis," Proc. IEEE Intl. Conf. on Neural Networks and Brain (ICNN&B), vol. 2, pp. 1202 – 1205, Beijing, China, Oct. 2005.

[17] S. Battiato, F. Rundo, F. Stanco, "ALZ: Adaptive Learning for Zooming Digital Images," Proc. IEEE Intl. Conf. on Consumer Electronics (ICCE), Las Vegas, USA, Jan. 2007.

[18] S. Dai, M. Han, Y. Wu and Y. Gong, "Bilateral Back-Projection for Single Image Super Resolution," in Proc. of IEEE Int. Conf. on Multimedia and Expo, Beijing, pp. 1039-1042, (July 2007).

[19] H. Ur and D. Gross, "Improved resolution from sub-pixel shifted pictures," CVGIP: Graphical Models and Image Processing, vol. 54, pp. 181-186, Mar. 1992.

[20] A. Papoulis, "Generalized sampling theorem," IEEE Trans. Circuits Syst. vol. 24, pp. 652-654, Nov. 1977.

[21] J.L. Brown, "Multi-channel sampling of low pass signals," IEEE Trans. Circuits Syst., vol. CAS- 28, pp. 101-106, Feb. 1981.

[22] R.C. Hardie, K.J. Barnard, and E.E. Armstrong, "Joint MAP registration and high-resolution image estimation using a sequence of undersampled images," IEEE Trans. Image Processing., vol. 6, pp. 1621-1633, Dec. 1997

[23] L. Landweber, “An iteration formula for Fredholm integral equations of the first kind,” Amer. J. Math. vol. 73, pp. 615-624, 1951

[24] T. Komatsu, T. Igarashi, K. Aizawa, and T. Saito, “Very high resolution imaging scheme with multiple different-aperture cameras,” Sinal Processing: Image Commun., vol. 5, pp. 511-526, Dec. 1993.

[25] M.S. Alam, J.G. Bognar, R.C. Hardie, and B.J. Yasuda, “Infrared image registration and high- resolution reconstruction using multiple translationally shifted aliased video frames,” IEEE Trans. Instrum. Meas., vol. 49, pp. 915-923, Oct. 2000.

[26] N.R. Shah and A. Zakhor, “Resolution enhancement of color video sequences,” IEEE Trans. Image Processing, vol. 8, pp. 879-885, June 1999.

[27] N. Nguyen and P. Milanfar “An efficient wavelet-based algorithm for image superresolution,” in Proc. Int. Conf. Image Processing, vol. 2, 2000, pp. 351-354.

[28] R.A. Roberts and C.T. Mullis, Digital Signal Processing, Addison-Wesley, 1987.

[29] A. M. Tekalp, M. K. Ozkan, and M. I. Sezan, “High-resolution image reconstruction from lower- resolution image sequences and space-varying image restoration,” in Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, San Francisco, CA, vol. 3, pp. 169–172, 1992.

[30] S.P. Kim, N.K. Bose, and H.M. Valenzuela, “Recursive reconstruction of high resolution image from noisy undersampled multiframes,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 38, pp. 1013-1027, June 1990.

[31] S.P. Kim and W.Y. Su, “Recursive high-resolution reconstruction of blurred multiframe images,” IEEE Trans. Image Processing, vol. 2, pp. 534-539, Oct. 1993.

[32] D. Keren, S. Peleg, and R. Brada, “Image sequence enhancement using subpixel displacements,” in Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, pp. 742– 746, June 1988.

[33] K. Aizawa, T. Komatsu, and T. Saito, “Acquisition of very high resolution images using stereo cameras,” Visual Communications and Image Processing Conf., 1991, Proc. SPIE, pp. 318–328, 1991.

[34] S. Peleg, D. Keren, and L. Schweitzer, “Improving image resolution by using subpixel motion,” Pattern Recognition Letters, vol. 5, no. 3, pp. 223–226, Mar 1987.

[35] V. Avrin and I. Dinstein, “Local Motion Estimation and Resolution Enhancement of Video Sequences”, in Proc. of the 14th IEEE Int. Conf. on Pattern Recognition, Washington, DC, USA, vol. 1, pp 539–541, Aug. 1998.

[36] D. Keren, S. Peleg, and R. Brada, “Image sequence enhancement using subpixel displacements,” in Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, pp. 742– 746, June 1988.

[37] M. Irani and S. Peleg, “Super Resolution From Image Sequences,” in Proc. of the 10th Int. Conf. on Pattern Recognition, Atlantic City, NJ, vol. 2, pp. 115–120, June 1990.

[38] M. Irani and S. Peleg, “Improving resolution by image registration,” CVGIP: Graphical Models and Image Processing, vol. 53, no. 3, pp. 231–239, May 1991.

[39] M. Irani and S. Peleg, “Motion analysis for image enhancement: Resolution, occlusion and transparency,” Journal of Visual Communications and Image Representation, vol. 4, no. 4, pp. 324–335, Dec. 1993.

[40] Suresh, K.V., Kumar, G.M., Rajagopalan, A.N.: Superresolution of license plates in real traffic videos. IEEE Trans. Intell. Trans. Syst. 8, 321–331 (2007)

[41] Narayanan, B., Hardie, R.C., Barner, K.E., Shao, M.: A computationally efficient super-resolution algorithm for video processing using partition filters. IEEE Trans. Circuits Syst. Video Technol. 17, 621–634 (2007)

[42] Ng, M.K., Shen, H.-F., Zhang, L.-P., Lam, E.: A total variation regularization based super- resolution reconstruction algorithm for digital video. EURASIP J. Appl. Signal Process. 2007(74585), 1–16 (2007)

[43] Patanavijit, V., Jitapunkul, S.: A Lorentzian stochastic estimation for a robust iterative multiframe super-resolution reconstruction with Lorentzian-Tikhonov regularization. EURASIP J. Appl. Signal Process. 2007(34821), 1–21 (2007)

[44] Elad, M., Feuer, A.: Superresolution restoration of an image sequence: adaptive filtering approach. IEEE Trans. Image Process. 8, 387–395 (1999)

[45] Elad, M., Feuer, A.: Super-resolution reconstruction of image sequences. IEEE Trans. Pattern Anal. Mach. Intell. 21, 817– 834 (1999)

[46] Farsiu, S., Elad, M., Milanfar, P.: Video-to-video dynamic super-resolution for grayscale and color sequences. EURASIP J. Appl. Signal Process. 2006 (61859), 1–15 (2006)

[47] D. Lowe, Distinctive Image features from scale invariant keypoints, International journal of Computer Vision, Vol. 60, pp. 91-110, 2004.

[48] H. Bay, T. Tuytelaars and L. Van Gool, SURF: Speeded Up Robust Features, Proc. European Conference on Computer Vision, Vol. 110, pp. 407-417, 2006.

[49] H. Bay, A. Ess, T. Tuytelaars, L. Gool, "Speeded-up robust features (SURF). Computer Vision and Image Understanding," vol. 110, no. 3,pp. 346-359, 2008.

[50] P.A. Viola and M.J. Jones, "Rapid object detection using a boosted cascade of simple features," In CVPR (1), pages 511-518, 2001.

[51] C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Images," Proc. IEEE Intl. Conf. on Computer Vision (ICCV), Bombay, India, Jan. 1998.

[52] http://r0k.us:80/graphics/kodak/

[53] http://www.cipr.rpi.edu/resource/stills/index.html

[54] http://www.cipr.rpi.edu/resource/stills/misc1.html

[55] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. on Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[56] R. Li, B. Zeng and M.L. Liou, "A new three-step search algorithm for block motion estimation," IEEE Trans. on Circuits and Systems for Video Technology, vol. 4, no. 4, pp 438–442, Aug. 1994.

[57] X. Jing and L.P. Chau, "An efficient three-step search algorithm for block motion estimation," IEEE Trans. on Multimedia, vol. 6, no. 3, pp 435–438, June 2004.

[58] Test video sequences: http://trace.eas.asu.edu/yuv/index.html.

[59] Test video sequences, http://media.xiph.org/video/derf/.

[60] http://code.google.com/p/zxing/

[61] Wang, X., Klette, R. and Rosenhahn, B. "Geometric and photometric correction of projected rectangular picture", Image and Vision Computing New Zealand, November 2005.

[62] A. Emilsson, “Indoor Navigation using an iPhone”, Linköpping University, 2010

[63] V. Pimenov, “Fast Image Matching with Visual Attention and SURF Descriptors,” In Proceedings of the 19th International Conference on Computer Graphics and Vision (GraphiCon'2009), Moscow, Russia Oct.2009