Video Transmission Robot Vision Motion-Noise Separation

A. Ritchie, J. Conradi, A. Prevot, and E. A. Yfantis
Image Processing Computer Graphics Intelligent Systems Lab. (ICIS)
School of Computer Science
Engineering College
University of Nevada, Las Vegas, 89154-4019
yfantis@cs.unlv.edu

Abstract In this research paper we examine the vision, between visit comparison, and transmission aspect of a robot used for remote medical hospital visits. The vision of the robot consists of two or more cameras, calibrated to resolve distances with centimeter to sub millimeter accuracy depending on the number of cameras used. The robot wirelessly transmits the lossy compressed ROY and lossy compressed the remaining of multiple images to a file server maintained in the hospital. The hospital file server transmits to the physician’s office real time or store and forward, and provides a comparison of the present images with the images of the previous visit. The software of our intelligent system provides a quantized comparison. In this research paper we address issues related to video capture for multiple cameras, noise reduction, de-interlacing, camera calibration, distance resolution, network programming, secure transmission, video compression, and video comparison.

Key Words: Network programming, De-Interlacing, Compression, Raid Arrays, Client, Server, Real-Time.

1. Introduction
As the population of many industrial countries ages the demand for health care increases while the number of physicians remains about the same or, in some places, has decreased. The need for telemedicine in many occasions provides an efficient way for health care professionals to increase the quality of care they provide to their patients. Robots, with the assistance of a nurse and a duplex network transmission, can enable physicians to visit with their hospital or home patients from the comfort of their office or clinic. Thus a physician can make patient visits in different hospitals without having to spend time traveling from her/his office and wasting valuable time on the road, thus avoiding traffic accidents, delays etc. The time saved can be invested in her/his patients care. Furthermore ambulances picking up patients from an accident scene, home, office or elsewhere, are able to transmit multiple camera video and other important information for real time evaluation and qualified medical professionals can provide advice. This knowledge could help paramedics treat patients efficiently, and emergency rooms to prepare for incoming patients so that they can provide the best of care. The importance of using quality cameras as opposed to capturing the video ad hoc with such methods like a cell phone is that one can control the video quality better, and also...
incorporate the video and audio as part of the software-hardware diagnostic intelligent system. Also, with multiple cameras it is easier to use de-interlacing [1], device noise reduction filters, resolve distances, compare and assess progress, and incorporate the rate of progress into a healing prediction model. The system consists of raid array servers, serving several robots in a hospital with many client computers at physician sites. The system also includes administration software that adds or deletes users, schedules robots, schedules robot visitation times, and adds or deletes robots. All network programming, and all other programming modules work in a parallel multithreaded environment. Although there are high quality cameras using high quality (R,G,B) CCDs producing progressive video at 30 frames per second, for many applications is advantageous to use cameras producing interlaced video, and then deinterlace the video and produce 60 frames per second video. Video is very demanding and requires a great deal of storage. Our system performs lossless and lossy compression on the same frame. Thus on the region of interest (ROY) we perform lossless compression [2], while on the rest of the frame we perform a lossy compression, using an algorithm we have developed, which is an extension of h.264, and conditional probability gradient prediction.

2.1 De-Interlacing

An image with resolution \(N\times M\) pixels can be thought of as a group of \(N\) signals each on having period of \((1/M)\). Thus to interpolate along a column we can use the Sinc function:

\[
f(x) = \sum_{n=0}^{N-1} f(n) \frac{\sin(\pi(x-n))}{\pi(x-n)}
\]

(1)

Thus,

\[
f(K) = \sum_{n=0}^{N-1} f(n) \cdot \left[ \frac{\sin(\pi(K-n))}{\pi(K-n)} \right]
\]

(2)

\[
= f(0) \sin \pi K + f(1) \frac{\sin \pi (K-1)}{\pi} + f(1) \frac{\sin \pi (K-2)}{\pi} + ... + f(N-1) \frac{\sin \pi (K-N+1)}{\pi}
\]

(K-2)

\[
+ f(K+1) \sin \pi + ... + f(N-1) \sin \pi \frac{K+1-N}{\pi}
\]

(3)

\[
f(K+\frac{1}{2}) = \sum_{n=0}^{N-1} f(n) \cdot \left[ \frac{\sin \pi(K+\frac{1}{2}-n)}{\pi(K+\frac{1}{2}-n)} \right]
\]

\[
= f(0) \sin \frac{\pi}{2}(2K+1) + f(1) \frac{\sin \pi (2K-3)}{(2K-3)/2} + ... + f(K) \frac{\sin \pi (2K-1)}{(2K-1)/2}
\]

(4)

\[
+ f(K+1) \sin \frac{\pi}{2} + ... + f(K+1) \sin \frac{\pi (K+1)}{2} + f(K+2) \sin \frac{\pi 3}{2} + f(K+3) \frac{\sin \pi 5}{\pi 2} + ... + f(K+4) \sin \frac{\pi 7}{2} + f(K+5) \sin \frac{\pi 9}{2} + ...
\]

In order to reduce the number of operations, if we need to double the resolution of the column, we can use 4 coefficients to the left and 4 to the right of the mid point to be interpolated. In order to do that the coefficients \((-2/\pi), (2/\pi), (-2/3\pi), (2/\pi), (2/\pi), (2/\pi), (2/\pi), (2/\pi),\)
\(-2/3\pi\), \(2/5\pi\), \(-2/7\pi\) have to be normalized.

The normalized coefficients are,
\[-2/7\pi a\), \((2/5\pi a\), \((-2/3\pi a\), \((2/\pi a\), \((-2/3\pi a\), \((2/5\pi a\), \((-2/7\pi a\)\] (5)

where,
\[a = 4 / \pi (1 - 1/3 + 1/5 - 1/7)\]
\[= (4 / \pi) * (76/105)\]

Thus the normalized coefficients are,
\[-105 / 14 *76\), \(105 / 10 *76\), \(-105 / 6 *76\), \((105 / 2 *76\), \((105 / 2 *76\), \(-105 / 6 *76\), \((105 / 10 *76\), \(-105 / 14 *76\)

So the coefficients are,
\[-0.0986842, 0.138158, -0.230263, 0.690789, 0.690789, -0.230263, 0.138158, -0.0986842\]

\[f(K+ ½ ) = -0.0986842 f(K-3)+ 0.138158 f(K-2)-0.230263 f(K-1)+ 0.690789 f(K)+ 0.690789 f(K+1)-0.230263 f(K+2)+ 0.138158 f(K+3) -0.0986842 f(K+4)\] (6)

Floating point operations are computationally expensive, an integer approximation to the above formula is:
\[f(K+½) = \frac{1}{100}[\frac{10f(k-3)+14f(k-2)-23f(k-1)+69f(k)+69f(k+1)-23f(k+2)+14f(k+3)-10f(k+4)]+30\]

The frequency response function of the above formula constitutes a bandpass filter which passes both low and high frequencies and produces sharp frames from fields, it increases the video quality and the video frequency from 30 frames per second to 60 frames per second. De-Interlacing is very helpful for camera calibration, noise-motion, separation, motion detection, motion estimation, and motion compensation, segmentation, feature vector formulation, classifier formulation, and pattern recognition.

2.2 Camera Calibration

Using the pinhole camera model [3-6], we can devise a mathematical model for the camera based on the principles of perspective projections.

Under this model, given a point \(\hat{X}\) in space, its projection onto the image plane \(\hat{p}\), is obtained by intersecting the ray joining the point \(\hat{X}\) and the center of projection \(C\) (camera center) with the image plane as shown in Figure 1.

![Perspective Projection](image)

**Figure 1:** Perspective Projection.

Using triangle similarities, one can derive the mapping between \(\hat{X}\) and \(\hat{p}\). In fact, if \(\hat{X} = (X, Y, Z)\), then \(\hat{p} = (f \frac{X}{Z}, f \frac{Y}{Z}, f)\) on the image plane, where \(f\) is the distance between the camera center \(C\) and the principal point \(O\) (which is the point of intersection between the principal axis and the image plane), and is referred to as the Focal Length. Thus, it follows that the mapping between world coordinates and image coordinates is given by:

\[
\begin{bmatrix}
X \\
Y \\
Z
\end{bmatrix}
\mapsto
\begin{bmatrix}
fX/Z \\
fY/Z
\end{bmatrix}.
\] (7)

To simplify the calculations, it is more convenient to express (1) using homogeneous coordinates. Indeed, the use of homogeneous coordinates
(also known as projective coordinates) allows for coalescence of a large number of transformations as well as all symmetries of the plane. Consequently, all transformations considered can be regarded as linear maps in the space of triplets \([x_1, x_2, x_3]^T\) and can thus be expressed in terms of matrix multiplications. Accordingly, (3.1) can be rewritten as

\[
\begin{bmatrix}
X \\
Y \\
Z \\
1
\end{bmatrix} = \begin{bmatrix}
fX \\
fY \\
fZ
\end{bmatrix} = \begin{bmatrix}
f & 0 & 0 & 0 \\
0 & f & 0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix} \begin{bmatrix}
X \\
Y \\
Z \\
1
\end{bmatrix}.
\]

In order to yield a more precise model for the camera, let us note that in Figure 1, we assume not only that the axes of the image coordinates space have equal scales but also that the origin of coordinates in the image plane is at the principal point. However, this may not be the case. In fact, in most CCD models, the scale in the axial directions may not be the same, so two new parameters need to be introduced to account for this omission. Thus, letting \(s_x\) and \(s_y\) be the scale factors in the \(x\) and \(y\) directions, respectively, and \((r_0, c_0)\) be the coordinates of the principal point, (2) becomes

\[
\begin{bmatrix}
X \\
Y \\
Z \\
1
\end{bmatrix} = \begin{bmatrix}
fs_x X + Z_0 \\
fs_y Y + Z_0 \\
0 \\
1
\end{bmatrix} = \begin{bmatrix}
f & 0 & r_0 & 0 \\
0 & fs_y & c_0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix} \begin{bmatrix}
X \\
Y \\
Z \\
1
\end{bmatrix}.
\]

Hence, if we let

\[
K = \begin{bmatrix}
fs_x & 0 & r_0 & 0 \\
0 & fs_y & c_0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix},
\]

we obtain

\[
\hat{p} = K \hat{X},
\]

where \(K\) is referred to as the \textit{Camera Calibration Matrix} and the parameters \(f, s_x, s_y, \text{ and } (r_0, c_0)\) are referred to as the \textit{Intrinsic Camera Parameters}.

Yet, our mathematical model is not accurate in view of the fact that, in Figure 3.1, we assume the points’ coordinates to be expressed with respect to the origin of the camera coordinate system, i.e., the camera center \(C\). However, in general, the coordinates of the points in space will be given with respect to a global Euclidean coordinate frame. Nevertheless, this can be easily overcome since the global coordinate system and the camera coordinate system are related by means of a pair of linear transformations, namely, a rotation and a translation (Figure 3.2). Let \(R\) and \(t\) be the rotation and translation that take the global coordinate system into the camera coordinate system, respectively. These are referred to as the \textit{Extrinsic Camera Parameters}.

\begin{figure}[h]
\centering
\includegraphics[width=0.8\textwidth]{camera_coordinates.png}
\caption{The Camera Coordinate System with respect to the Global (World) Coordinate System.}
\end{figure}
In consequence, given $\mathbf{X}_w$, a point in space with coordinates in the global coordinate system, its coordinates with respect to the camera coordinate system, denoted by $\mathbf{X}_c$, are then given by

$$\mathbf{X}_c = \mathbf{R} \mathbf{X}_w + \mathbf{t}. \quad (11)$$

Therefore, combining (3.4) and (3.5), it follows that

$$\mathbf{p} = K \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \mathbf{X}_w, \quad (12)$$

where $\mathbf{X}_w$ denotes the homogeneous coordinates of the three-dimensional point $\mathbf{X}_w$. That is,

$$\mathbf{p} = P \mathbf{X}_w, \quad (13)$$

where $P = K \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}$ is called the Projection Matrix.

In what follows,

$$P^1 = K^1 \begin{bmatrix} \mathbf{R}^1 & \mathbf{t}^1 \end{bmatrix}$$

and

$$P^2 = K^2 \begin{bmatrix} \mathbf{R}^2 & \mathbf{t}^2 \end{bmatrix} \quad (14)$$

will denote the projection matrices for cameras 1 and 2, respectively.

A familiarity with the concept of epipolar geometry will be beneficial to the reader since it is of importance in our reconstruction algorithm. Thus, we describe it in the next section. The epipolar geometry is the intrinsic projective geometry between two views [8, 9]. Among the most important characteristics of the epipolar geometry between two views are the facts that it depends only on the cameras’ internal parameters and relative pose and that it is independent of the scene structure. In what follows, a concise description of the most relevant concepts of the epipolar geometry will be presented; for in Chapter 5, it will be necessary to understand the relation between the projections, onto two different views, of a point in space.\[\mathbf{X}_w \in \mathbb{R}^3, \quad \mathbf{p}_1, \mathbf{p}_2\] be its projections on Image 1 (corresponding to camera 1) and Image 2 (corresponding to camera 2), respectively. The goal is to find a relation between $\mathbf{p}_1$ and $\mathbf{p}_2$.

Clearly, $\mathbf{X}_w, \mathbf{p}_1, \mathbf{p}_2$, and the camera centers $\mathbf{c}_1$ and $\mathbf{c}_2$ are coplanar (Figure 3); $\mathbf{X}_w$ lies in the plane $\pi$ which is referred to as the epipolar plane.

![Figure 3: 3D Space Point.](image)

Moreover, given that $\mathbf{p}_1$ and $\mathbf{p}_2$ are the projections of $\mathbf{X}_w$, the rays emanating from the respective camera centers and passing through $\mathbf{p}_1$ and $\mathbf{p}_2$ not only intersect at $\mathbf{X}_w$, but they also lie in $\pi$.

We can also observe, from Figure 3, that the plane $\pi$ can be uniquely determined by considering only one of the projection rays (provided one knows the coordinates of the 3D point and one of its projections) and the line segment connecting the camera centers, which is referred to as the baseline. Let $L_i$ be the projection ray determined by $\mathbf{p}_i$ and $\mathbf{X}_w$. $L_i$ is projected onto Image 2...
as the line $\ell_{12}$. Moreover, the image of $\hat{X}_W$ in the view from camera 2, i.e $\hat{p}_2$, must lie on $\ell_{12}$. Hence, if one assumes that $\hat{p}_1$ is known, the search for its matching point on the second view is then limited to points lying on the epipolar line $\ell_{12}$. Clearly, the same argument is valid for the reciprocal case, i.e., knowing $\hat{p}_2$, its matching point in the first view can be found via a search along the epipolar line $\ell_{21}$.

The images are compressed. Specifically, the region of interest is compressed losslessly while the remaining frame is compressed lossy. In addition to compression we provide Cryptography based on our own extension of AES and also we provide forward error correction to ensure that the images have not been corrupted during their wireless and then possibly wired channel transmission. The Hospital raid array server consists of over 100 terabytes and it is a software controlled server, that hosts video from many patients and a database associated patients with Physicians, robots used in each visit, nurses escorting the robots, time of the visit, duration of the visit, and commend from the physician related to this visit. Each physician has access to her/his patient video and information and can download a copy to her/his computer dedicated to this activity.

3. Conclusion
In this research paper we examined the vision, comparison, and transmission aspect of a robot used for remote medical hospital visits. The vision of the robot consists of two cameras or more cameras, calibrated to resolve distances with centimeter accuracy. The robot transmits wirelessly the multiple images to a file server maintained in the hospital. The hospital file server transmits to the physician’s office real time, or store and forward, and provides a comparison of the present images with the images of the previous visit. The physician makes a visual comparison and the software of our intelligent system makes a comparison and provides intelligent and quantized progress information. In this research paper we addressed issues related to video capture for
multiple cameras, noise, de-interlacing, camera calibration, distance resolution, network programming, secure transmission, and video comparison.

References: