1. Introduction

Tele-immersion may be the next major development in information technology. Using tele-immersion, you can visit an individual across the world without stepping a foot outside.

1.1 What is tele-immersion?

Tele-Immersion is a new medium that enables a user to share a virtual space with remote participants. The user is immersed in a 3D world that is transmitted from a remote site. This medium for human interaction, enabled by digital technology, approximates the illusion that a person is in the same physical space as others, even though they may be thousands of miles distant. It combines the display and interaction techniques of virtual reality with new computer-vision technologies. Thus with the aid of this new technology, users at geographically distributed sites can collaborate in real time in a shared, simulated, hybrid environment submerging in one another's presence and feel as if they are sharing the same physical space.

It is the ultimate synthesis of media technologies:

- 3D environment scanning,
- projective and display technologies,
- tracking technologies,
- audio technologies,
- robotics and haptics,

and powerful networking. The considerable requirements for tele-immersion system, make it one of the most challenging net applications.

In a tele-immersive environment computers recognize the presence and movements of individuals and objects, track those individuals and images, and then permit them to be projected in realistic, multiple, geographically distributed immersive environments on stereo-immersive surfaces. This requires sampling and resynthesis of the physical environment as well as the users' faces and bodies, which is a new challenge that will move the range of emerging technologies, such as scene depth extraction and warp rendering, to the next level.

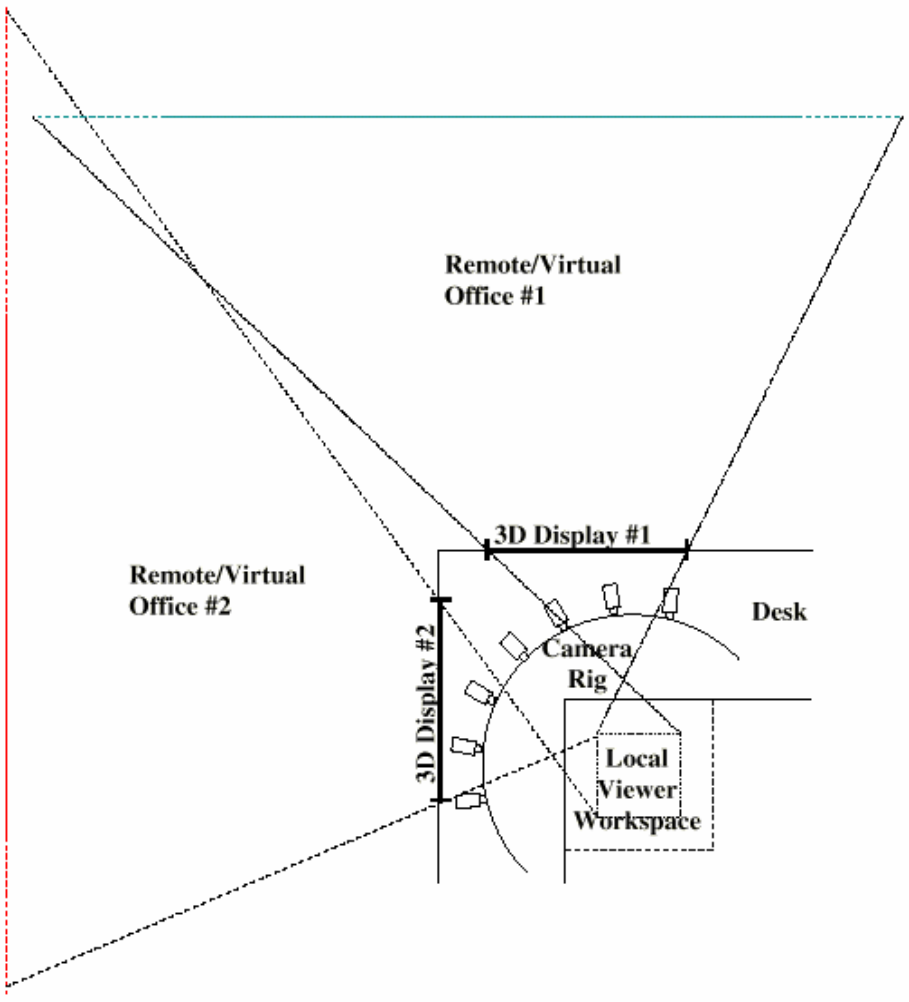
Tele-immersive environments will therefore facilitate not only interaction between users themselves but also between users and computer generated models and simulations. This will require expanding the boundaries of computer vision, tracking, display, and rendering technologies. As a result, all of this will enable users to achieve a compelling experience and it will lay the groundwork for a higher degree of their inclusion into the entire system

Tele-immersive systems have potential to significantly change educational, scientific and manufacturing paradigms. They will show their full strength in the systems where having 3D reconstructed ‘real’ objects coupled with 3D virtual objects is crucial for the successful fulfillment of the tasks. It may also be the case that some tasks would not be possible to complete without having such combination of sensory information. There are several applications that will profit from tele-immersive systems. Collaborative mechanical CAD applications as well as different medical applications are two that will benefit significantly.

Tele-immersion may sound like conventional video conferencing. But it is much more. Where video conferencing delivers flat images to a screen, tele-immersion recreates an entire remote environment. Although not so, tele-immersion may seem like another kind of virtual reality. Virtual reality allows people to move around in a pre-programmed representation of a 3D environment, whereas tele-immersion is measuring the real world and conveying the results to the sensory system.

2. System Overview And Algorithms

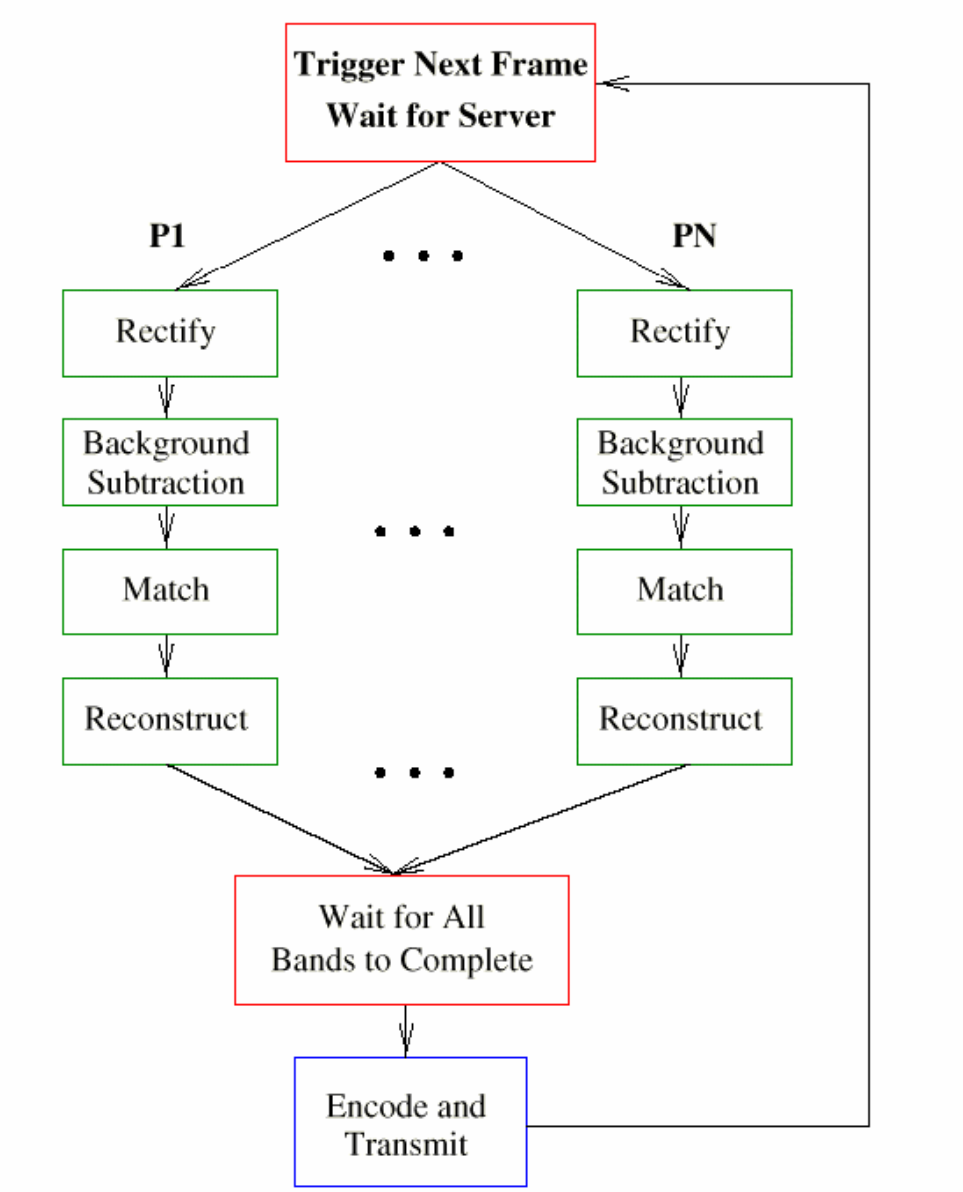
A tele-immersion telecubicle is designed both to acquire a 3D model of the local user and environment for rendering and interaction at remote sites, and to provide an immersive experience for the local user via head tracking and stereoscopic display projected on large scale view screens. A typical setup can be depicted as follows.



The user moves freely in a 1m workspace at his desk. Remote users are rendered on 90cm X 120 cm screens by projector pairs. The user wears lightweight polarized glasses and a head-tracker to drive the stereo display function. A cluster of 7 firewire cameras are arranged on an arc at 15^{o} separation to 'surround' the user and prevent any break presence due to hard edge where the reconstruction stops. These cameras are used to calculate binocular or trinocular stereo depth maps from overlapping pairs or triples. The technical obstacle to the combining of camera views, is that each reconstruction is performed on a separate computer which adds to the overhead of the system.

Both responsiveness and quality of depth data are critical for immersive applications. In order to improve the frame rate of the system a number of techniques can be applied to reduce the weight of calculation, particularly in the expensive correlation matching required to generate dense depth maps. The simplest technique for the developer is , of course, to purchase more and faster computers. Our system uses rectification, background subtraction, correlation matching and median filtering to balance quality and speed in our reconstructions. A general parallel structure of the system can be illustrated as follows.

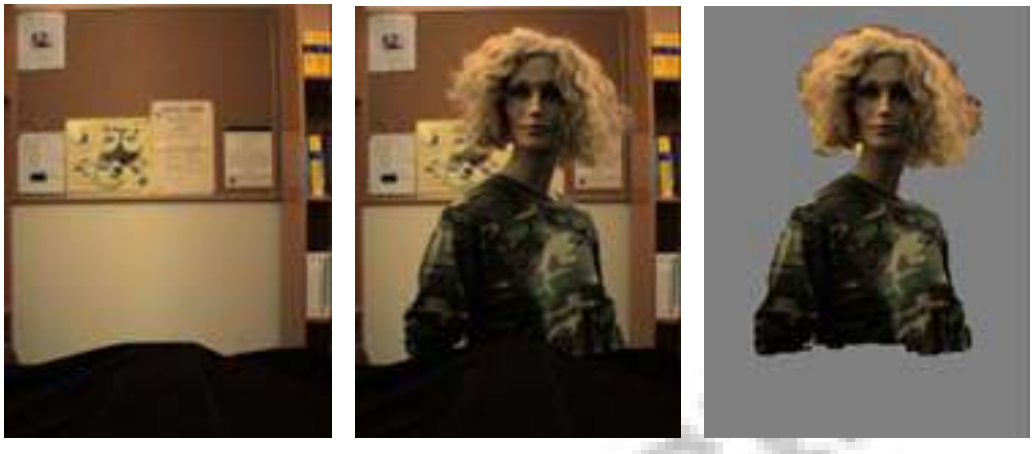
One of the servers acts as a trigger server for the firewire acquisition. When all of the reconstructors are ready for the next frame the trigger server triggers all of the cameras simultaneously. Each computer grabs the image from 1 or 2 cameras and transmits and receives the images needed by its neighbours and itself. Within each quad machine, the images are divided into 4 equal bands and each processor is devoted to a particular band. The thread for each processor rectifies, background subtracts, matches and reconstructs points in its band of the image. When all the processors have completed processing the texture and depth map are transmitted via TCP/IP to a remote renderer. It takes about 3 Mbits per view per frame.



2.1 Background Subtraction

It is expected that the workspace will contain a person in the foreground interacting with remote users and a background scene which will remain more or less constant for the duration of a session. To obtain the speed and quality of depth points the application requires, the background scene can be reconstructed in advance of the

session and transmitted once to the remote sites. While the user moves into the foreground during a session, the static parts are segmented out.



Background Image Foreground Image Subtracted Result

To further optimize calculation we compute the foreground mask for both images of the reference pair. In this way fore-ground pixels are only matched against foreground pixels, not background pixels.

A sequence of N (2 or more) background images B_i are acquired in advance of each session. From this set a pixelwise average background image can be calculated as: $B_{avg} = \Sigma_i B_i / N$. Then we can compute the average poxelwise difference between B_{avg} and B_i as : $D = \Sigma_i (B_{avg} - B_i) / N$. During a tele-immersion session each primary image I is subtracted from the static mean background $I_D = B_{avg} – I$, a binary image is formed via the comparison $I_B = I_D > T \times D$, where T is a configurable threshold. These thresholded differences are generally quite noisy. A series of erosions and dilations are performed on the binary image in order to sharpen the background mask.

2.2 Bi- and Trinocular Matching

Different types of correlation correspondence techniques can be used for creating dense stereo depth maps. But consideration is to be given to the aspects of

speed and quality. Sum of Absolute Differences (SAD) has been found advantageous because of the speed provided by hardware specific operations. Also, Modified Normalized Cross Correlation (MNCC) can be employed as it produces superior depth maps in the binocular case. Trinocular SAD and MNCC may also be used.

2.2.1 Correlation Methods

The reconstruction algorithm begins by grabbing images from 2 or 3 strongly calibrated cameras. The system rectifies the images so that their epipolar lines lie along the horizontal image rows to reduce the search space for correspondences, and so that corresponding points lie on the image lines.

The calculation of SAD as a correlation metric is facilitated on Intel/MMX machines by an assembler operation which can calculate the sum of absolute differences between two registers. In general, the SAD calculation is:

$$corr_{SAD}(I_L,I_R) = \Sigma_W |I_L-I_R|$$

for a window W in rectified images I_L and I_R. The disparity *d* determines the relative window position in the right and left images.

A better correspondence metric is modified normalized cross correlation (MNCC).

$$corr_{MNCC}(I_L,I_R)= 2cov(I_L,I_R) / (\sigma^2.I_L+\sigma^2.I_R)$$

where I_L and I_R are the left and right rectified images over the selected correlation windows.

For each pixel (*u,v*) in the left image, the metrics above produce a correlation profile *c*(*u,v,d*) where disparity *d* ranges over acceptable integer values. Selected matches are maxima (for MNCC) or minima (for SAD) in this profile.

2.2.2 Non-parallel Trinocular Configurations

The trinocular epipolar constraint is a well known technique to refine or verify correspondences and improve the quality of stereo range data. It is based on the fact that for a hypothesized match [*u,,v,d*] in a pair of images, there is a unique location we can predict in the third camera image where we expect to find evidence of the same world point. A hypothesis is correct if the epipolar lines for the original point [*u,v*] and the hypothesized match [*u – d, v*], intersect in the third camera image. The most common scheme for exploiting this constraint is to arrange the camera triple in a right angle, allowing matching along the rows and columns of the reference image. The telecubicle configuration already explained does not allow us to arrange or rectify triples of camera image planes such that they are coplanar, and therefore it is more expensive to exploit the trinocular constraint.

The sum of correlation values are treated with respect to true depth rather than disparity. Essentially we treat the camera triple (L,C,R) as two independent stereo pairs (L,C_L) and (C_R , R), using the (L,C_L) pair to verify matches in the right reference pair (C_R , R). This method involves usage of pre-computed correlation images for ranges of disparity in the left camera pair, then the computed correlation for each tested () was added to that corresponding (). This results in large correlation lookup tables for the left image pair.

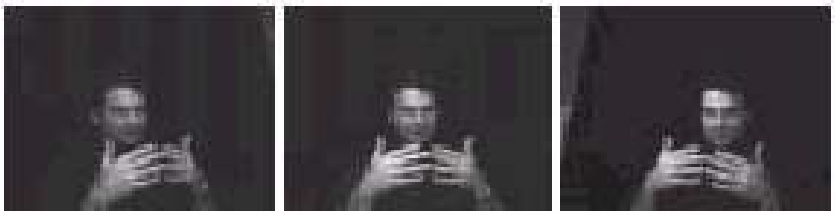
A second method is developed to avoid the large lookup tables by independently finding the best N extrema in the correlation surfaces for both image pairs. These sorted hypotheses were then cross checked to determine whether a common depth point gave rise to the scores for any pair. Valid hypothesis pairs with the best score were retained. This method required less lookup table space, but had considerable added overhead to maintain the sorted hypotheses.

3. Performance and Results

Methods exploiting SAD were faster than MNCC based implementations. All implementations ran on a quad PIII 550 MHz server in 1 second or less, including image acquisition and transfer and transmission of reconstructions to the renderer. Timings for the various systems are presented in the table:

Step	SAD	MNCC	Tri-SAD	Tri-MNCC
Rectify	49	50	49	48
Background	18	18	18	18
Matching	182	261	390	791
Reconstruct	6	6	7	6
Total	446 ms	520 ms	662 ms	1067 ms
fps	2.2	1.9	1.5	0.9

For tele-immersion the quality and density of depth points are most important. Although computation times are greater, the high quality of trinocular depth maps makes them a desirable alternative to faster but noisier SAD images. Figures below illustrate a trinocular triple and the resulting rendered depth maps for binocular MNCC (right pair) and trinocular MNCC respectively.



Trinocular triple

Rendered reconstructions, profile view (a) Binocular MNCC (b)Trinocular MNCC

The improvement in depth map from use of the trinocular constraint is evident in the reduction of noise speckle and refinement in detail.

An added challenge with the seven camera cluster is the combination of multiple reconstructions into a single rendered view. Figure below shows a full set of camera views for a single frame in the current telecubicle camera cluster. From this image set, 5 reconstructed views are calculated for overlapping triples. The second figure below shows a profile rotation of the total set of 164000 depth points calculated using trinocular MNCC for the frame in the first figure.

Seven camera views.

Five trinocular reconstructions combined
and rendered, rotated view.

4. Motion Based Enhancements

The dominant cost in stereo reconstruction is that of the correlation match itself, in general proportional to N x M x D for images of size N x M and D tested disparity values. By using background subtraction in our application we have reduced the number of pixels considered by one half to one third of the total N x M. To reduce the matching costs further, the number of disparities, D, has to be reduced.

A further observation regarding online stereo reconstruction is that for high frame rates there will be considerable similarity between successive images. This temporal coherence can be exploited in order to further optimize out online calculations. For this, we can perform simple segmentation of the image, based on finding regions of the image which contain only a narrow range of disparity values. Using a per region optical flow calculation we can estimate the location of the region in future frames and bound its disparity search range Di.

A method for integrating disparity segmentation and optical flow can be summarized in the following steps.

Step 1: Bootstrap by calculating a full disparity map for the first stereo pair of the sequence.

Step 2: Use flood-fill to segment the disparity map into rectangular windows containing a narrow range of disparities.

Step 3: Calculate optical flow per window for left and right smoothed , rectified image sequences of intervening frames.

Step 4: Adjust disparity window positions and disparity ranges according to estimated flow.

Step 5: Search windows for correspondence using assigned disparity range, selecting ‘best’ correlation value over all windows and disparities associated with each pixel location.

Step 6: Go to Step 2.

Most time critical systems using correlation matching benefit from this approach as long as the expense of propagating the windows via optical flow calculations is less than the resulting savings over the full image/ full disparity match calculation.

Restricting the change in disparity per window essentially divides the underlying surfaces into patches where depth is nearly constant. A threshold is used on the maximum absolute difference in disparity as the constraint defining regions, and regions are allowed to overlap. Only rectangular image windows are maintained rather than a convex hull or more complicated structure, because it is generally faster to apply operations to a larger rectangular window than to manage a more complicated region structure. Regions are extracted using flood fill or seed fill, a simple polygon filling algorithm from computer graphics.

Optical flow calculations approximate the motion field of objects moving relative to the cameras, based on the image brightness constancy equation: $I_x v_x + I_y v_y + I_t = 0$, where I is the image brightness and I_x, I_y and I_t are the partial derivatives of I with respect to x,y and t, and $v = [\ v_x,\ v_y\ ]$ is the image velocity. A standard local weighted least square algorithm can be used to calculate the values for v based on minimizing :

$$e = \sum_{W_i} (I_x v_x + I_y v_y + I_t)^2$$

In the case of our disparity windows, each window can be of arbitrary size but will have relatively few disparities to check. Because our images are rectified to align the epipolar lines with the scanlines, the windows will have the same y coordinated in the right and left images. Given the disparity range, we can extract the desired window from the right image given by $x_r = x_l - d$.

It has been demonstrated experimentally that the window based reconstructions compare favourably to those generated by correlation over the full image, even after several frames of propagation via estimated optical flow. The observed mean differences in computed disparities were less than one pixel and the maximum standard deviation was 4.4 pixels.

5. Overcoming Technical Issues

There are quite a few difficulties involved. The first is, how do you sense a remote place in real time fast enough and with the kind of quality so you can re-render it and make it look good? There you have a mixture of vision problems, graphics problems, and networking problems all in one bundle. Then beyond that, how do you create a physical viewing configuration that supports the illusion of reality and keeps it almost real?

Starting with the scene acquisition of a complete three-dimensional representation independent of any one perspective, vision techniques were employed, using a "sea" of multiple video cameras. For the best trade-off of quality versus performance, overlapping trios of video cameras-with more redundancy of scene information allowing fuller coverage of visual surfaces than with pairs were employed. In the advanced teleconferencing application, seven video cameras are arranged in a 120-degree arc in front of each subject, with the cameras sampled in overlapping triads-the optimumarray, given network constraints. The goal is for higher resolution images with a 60-camera array that can be used in medical tele-mentoring.

Among the graphics problems to be overcome is that of surface ambiguities that the human brain resolves effortlessly but computers have difficulty parsing. Under normal room illumination, such as that from overhead fluorescent bulbs or desktop incandescent lamps, a bare wall or a shiny surface will not display any surface textures at all, which will confound thepattern-recognitionsoftware.

In an attempt to accurately register featureless walls, and shiny objects for that matter, the team is exploring a technique developed at UNC called "imperceptible structured light." With this technology, along with a room's existing lighting, a scene is lit by what appear to be additional normal spotlights, but embedded with in this illumination are structure-monochromatic geometric patterns. The patterns are dithered to be imperceptible to people, but a synchronized video camera can pick up these visual calibration patterns so that ambiguities in the shape, color, and reflectivity of objects

(such as screens, blank walls, doorways, and even a person's forehead) are eliminated.

The stated goal of real-time scene acquisition and transmission accentuates graphics and networking challenges. For example, sampling rates vary with the complexity and movement in a scene as pattern-recognition algorithms model visual content captured by the video camera arrays; a chosen algorithm's accuracy must be weighed against any lag time it may introduce. To cut down on computational lag, some optimization is performed, such as segmenting a scene so more resources can be devoted to accurately capturing human facial features. Such feature recognition is the underlying technology for Eyematic, a Los Angeles-based start-up.

Once the visual information is received at a remote location, it is re-rendered as computer-generated people and sets by a computer specialized for this task. Some problems remain in finding the best techniques for depicting different kinds of objects; for instance, hair renders better as a point cloud, whereas skin renders better as polygons with textures.

To enhance the illusion of reality, stereoscopy is essential for depth perception. The display currently used to achieve this in tele-immersion consists of a pair of front projectors for each transmitted scene, with polarized-lens glasses worn by participants to separate reconstituted right- and left-eye views.

Accurately establishing a teleconference participant's point of view for proper re-rendering of remote scenes now requires another awkward apparatus: a UNC-developed headtracker based on 3rdTech's HiBall tracker, which sits on the user's head like a silver salt shaker. It establishes the viewer's physical orientation by sensing position relative to infrared light-emitting diodes embedded in the ceiling.

It is expected that the cost reductions implicit in Moore's Law to apply to video chips as well, ultimately resulting in position tracking by visual sensors imprinted like wallpaper in a room, thus eliminating cumbersome headgear.

Further limiting the real-time capabilities of tele-immersion are the physics of light itself. Photons passing through fiber optics travel far more slowly than the universal speed of light, causing perceptible transmission delays over long distances. For a wide variety of applications, a lag of 30 to 50 milliseconds is what is acceptable to viewers. In some cases, critical elements like head and hand movements can be tweaked with predictive algorithms that in effect accelerate scene capture beyond the moment in time, inferring the next position of a hand gesture or head movement, but these can only do so much. Perhaps by harnessing super computers, any lag could be eliminated. For now, the maximum distance for tele-immersion to be effective is roughly the width of the United States of America.

The problem is, today's internet can't ship data nearly fast enough. To look anything like reality, tele-immersion will have to be able to move mountains of data – spatial and visual information about people and their environments – across the internet in a trice. Today's 56kbps connections cannot do that. Even the bare-bones demonstrations of tele-immersion that has been undertaken until now needed 60Mbps. High quality tele-immersion will require even more – around 1.2Gbps. Fortunately the kind of capacity that is needed is on the way in the form of internet2, a consortium of American universities, government agencies, private companies and international organizations that is trying to recreate the collaborative spirit of the early internet. It is predicted to be a unique test bed for the future internet applications, including teleimmersion. Also, teleimmersion will require supercomputers to perform the trillions of calculations that are needed to portray environments in 3D. This kind of computer power would have to be on the tap over the internet. Something like that is on the way, too, in the form of a network called the Grid.

6. Potential Applications

Tele-immersive systems have potential to significantly change educational, scientific and manufacturing paradigms. They will show their full strength in the systems where having 3D reconstructed 'real' objects coupled with 3D virtual objects is crucial for the successful fulfillment of the tasks. It may also be the case that some tasks would not be possible to complete without having such combination of sensory information.

There are several applications that will profit from tele-immersive systems. Collaborative mechanical CAD applications as well as different medical applications are two that will benefit significantly. For example, a group of designers will be able to collaborate from remote sites in an interactive design process. They will be able to manipulate a virtual model starting from the conceptual design, review and discuss the design at each stage, perform desired evaluation and simulation, and even finish off the cycle with the production of the concrete part on the milling machines.

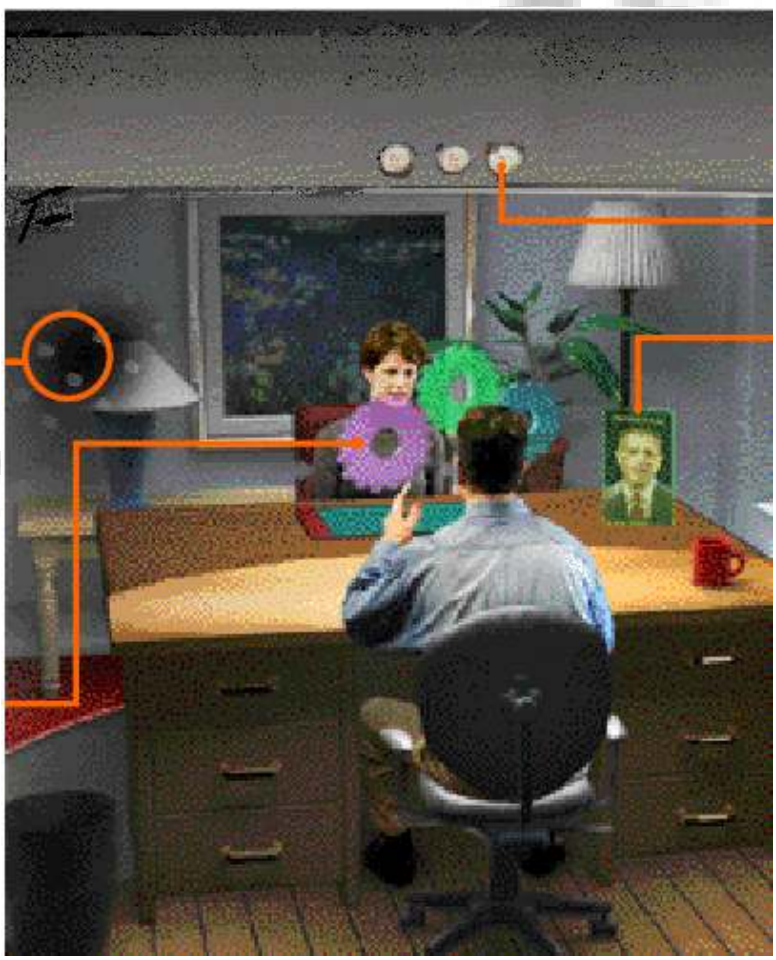
In the case of medical applications such as tele-radiology and urgent diagnostics, the availability of such technologies in the places that are physically inaccessible to specialists could potentially save the lives. Off-shore ships and oil rigs are good examples of such environments.

Tele-immersion will enable researchers to collaborate in fields such as architecture, medicine, astrophysics and aeroplane design. It allows widely separated people to share a complex virtual experience.

Tele-immersion is not just a research tool. Restaurants fitted with tele-immersion booths will enable people far away from home to have dinner with their family. The gaming industry is another potential user. Players could tele-immerse themselves in a virtual reality environment, chasing monsters or firing phasers at each other.

7. The Future

Researchers aim to make tele-immersion more natural, by jettisoning the headgear and glasses altogether. It is expected that a person should be able to experience tele-immersion by just entering a tele-cubicle. One possibility is to use a screen that transmits different information to each eye, using swiveling pixels that track either left or right eye. Another idea is to turn the entire tele-immerion room onto a screen. Walls, tables, curtains, even floors could be coated with special light sensitive material. Camera would photograph the surfaces, computers would calculate their shapes in 3D, and projectors would shine pre-warped images, making it seem as if they filled the room.



(1)Imperceptible structured light. (2)Sea of cameras .
(3)Virtual mirror (4)Shared simulation objects

The above picture shows a tele-cubicle from the future. The virtual objects can be pointed at by using virtual laser pointers. Gone will be the days of the seven prominent cameras facing the user. Instead, cameras will be placed somewhere in the tele-cubicle where it is less prominent. It is expected that there will be a sea of around 50 to 60 cameras in a tele-cubicle to provide a perfect tele-immersive experience.

Imperceptible Structured Lights are going to be a standard part of tomorrow's tele-cubicle. These help in resolving surface ambiguities due to which the computer finds it difficult to recognize what a surface or object is. The virtual mirror enables a user to see how he himself is being viewed by other participants. All users in a particular session can manipulate the shared simulation objects.

In future, it will be possible to manipulate virtual objects. The first prototype of Virtual Reality Mail System has already been developed. In VR-mail, users make a recording by speaking and gesturing. The audio and gestures are captured and saved in a format that allows a synchronized playback at a later time. This recording can then be sent to another user in the Virtual Environment (VE). When the recipient of the message enters the VE, he or she will find a VR-mail message waiting for him or her. The recipient may then play back the message. As in a traditional e-mail system, the recipient is then able to respond to the original sender of the VR-mail. In future, this idea can be extended to Tele-Immersion as well.

Researchers from the University of Illinois at Chicago are aiming to merge tele-immersion and virtual reality together whereby people can share a virtual environment and each is visible to the others as a computer simulated entity or "avatar". People could choose the way they look in a tele-immersion session – from changing their hair colour to looking like a film star.

8. Conclusion

Tele-immersion techniques can be viewed as the building blocks of the office of tomorrow, where several users from across the country will be able to collaborate as if they're all in the same room. Scaling up, transmissions could incorporate larger scenes, like news conferences, ballet performances, or sports events. With mobile rather than stationary camera arrays, viewers could establish tele-presence in remote or hazardous situations.

Far from just a validating application for the next-generation Internet, tele-immerion is expected to fundamentally change how we view real and virtual worlds.

REFERENCES

9. Who all are involved ?

Tele-Immersion is still under development, under the National Tele-Immersion Initiative (NTII). The main participants of NTII include

- Advanced Network and Services, Armonk, New York
- Brown University, Rhode Island
- University of North Carolina, Chapel Hill, North Carolina
- University of Pennsylvania, Philadelphia

In the past, several other organizations were involved with Tele-Immersion and they include

- Naval Post Graduate School, California
- Carnegie Mellon University, Pennsylvania
- Columbia University, New York
- University of Illinois, Chicago and
  University of Southern California, California.