ICAT 2005
December 5-8, Christchurch, New Zealand

Augmented Telexistence in Smart Space

Ig-Jae Kim, Sang-Yup Lee, Sang Chul Ahn and Hyoung-Gon Kim
Imaging Media Research Center, KIST
39-1 Hawolgok-dong, Seongbuk-gu, Seoul, Korea
drjay@kist.re.kr

Abstract

In this paper, we present the new environment for telexistence using our smart space which we have built for interactive space. We've developed several component modules for telexistence, such as display module for large scale display system, interaction module for multimodal interaction, video module for streaming and rendering module for augmented composition in our smart space. Those systems are well-integrated and allow for augmented telexistence functionality which is more natural and also give a beyond face-to-face communication between remote users. Based on our proposed environment, we can create new possibilities for immersive and interactive communication across the space.

Key words: Telexistence, Multimodal Interaction, Tiled Display, Smart Space, Live Avatar

1. Introduction

Telexistence is the concept of overcoming limitations of time and space by presenting a sensation of existence in a virtual sense. Technology based on this concept enables users to meet and talk to one another as if they share the same space and time, even if they are far apart from each other[1]. Telexistence can be divided into two categories, telexistence in the real world and telexistence in a virtual environment. Telexistence in the real world, which actually exists at a distance and is connected via a robot to the user's location, and telexistence in a virtual environment, which does not actually exist but is created by a computer. The former can be called "transmitted reality," while the latter is "synthesized reality." The synthesized reality can be classified into two, i.e., a virtual environment as a model of the real world and a virtual environment of an imaginary world. Combination of transmitted reality and synthesized reality, which is called mixed reality, is also possible and has great importance in real applications. This is an augmented telexistence to clarify the importance of harmonic combination of real and virtual environments[2]. In our smart space environment, we can combine the transmitted reality and synthesized reality efficiently based on our component modules for telexistence described in figure 1. Video module plays an important part for transmitting reality from remote world and also provides the solution for vision based approaches. Interaction module functions as an I/O for remote users to interact. For multimodal interaction, we've built smart floor and camera array for tracking and detecting user. In rendering module, we integrated several functionalities such as composition functionality which compose virtual environment as a model of the real world with synthetic objects, live avatar functionality which provides coexistence with remote users. And display module provides large scale display and seamless display for immersiveness.

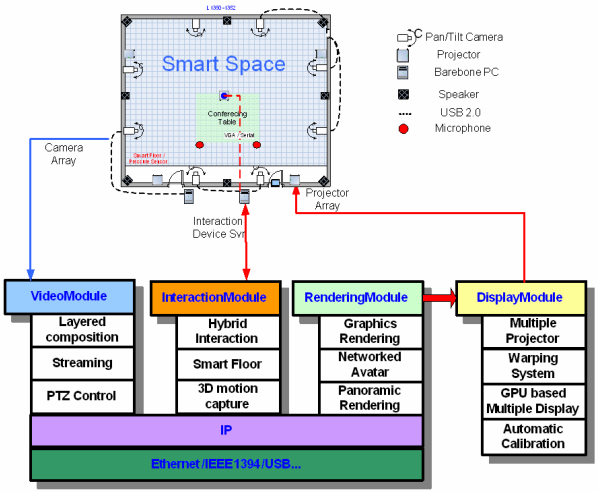


Fig. 1 Component architecture in our smart space

2. Smart Space

In our smart space, we integrated several systems, as mentioned before, such as display module, interaction module, rendering module and video module for telexistence. We'll introduce the details for those modules in the following subsection.

2.1 Display Module

We have developed a simple and efficient design for network-based tiled display system to solve the problems that previous approaches which depend on a special platform and library to make contents of tiled display system. We have developed the proposed system using GPU based real-time warping with camera based automatic feature extraction method, blending overlapping area and network-based streaming. Finally, different from the existing systems which used under the specific application environment, we can use any commercial application without modification with high resolution. In order to build a

tiled display, we have to go through several steps, that is, streaming, two pass warping, and blending process.

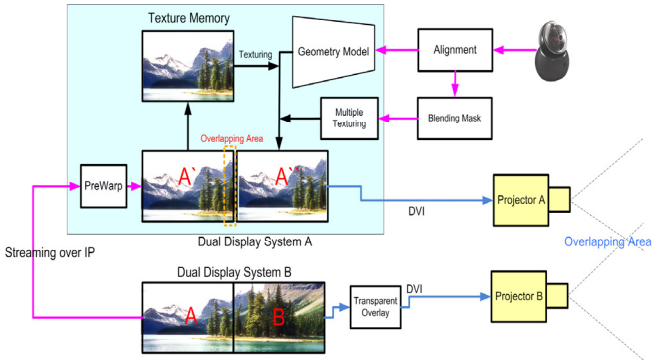


Fig. 2 Total flow of our display module

2.1.1 Streaming

In this stage, we transmit the reference view which includes overlapping area of dual display system B to the pre-warping module in the dual display system A as shown in figure 2. For this process, we use transmission module in Virtual Network Computing[3] and modify it to increase the performance of transmission. After transmitting the reference view of dual display system B to the prewarp module of system A, we make a pre-warping process to fit the overlapping area given by two projectors.

2.1.2 Two Pass Warping

After transmitting the reference view of dual display system B to the prewarp module of system A, we make a two pass warping process, that is, pre-warping and post-warping process. In pre-warping process, a transmitted image is warped to fit the overlapping area given by two projectors. We extract warping parameters, that is, a translation term for similarity transform, through the alignment process using camera based automatic feature extraction technique.

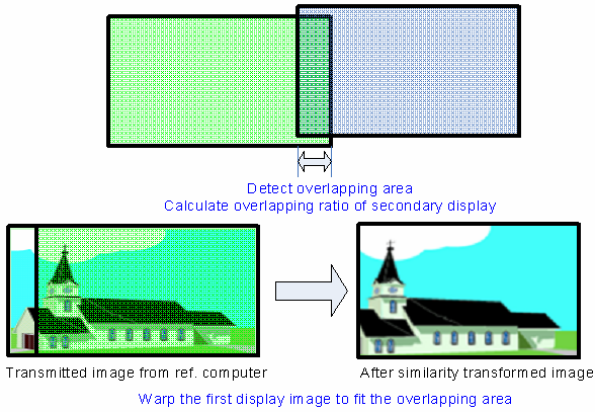


Fig. 3 Prewarp process

And then, we need image warping to stitch two images and to make them into an image with homography given by alignment process. This work is done in post-warping process. Since the realtime image warping requires large computational complexity and high speed image processing, previous approaches used special hardware to support those kinds of work. However, there has been a great improvement in the performance of GPUs, we can utilize the power of GPUs to handle the realtime warping process. For this reason, we decided to use DirectX to control GPUs[4]. In post-warping process, we use dual display functionality of graphic card with prewarped image. This process is described in green area of figure 1. We display the prewarped image on the first monitor of dual monitor and capture the displayed image and load it to the texture memory. After mapping it to the gemetry model, we can warp the image by transforming the geometry vertices. We finally display the warped image on the second monitor(in this case, projector A).

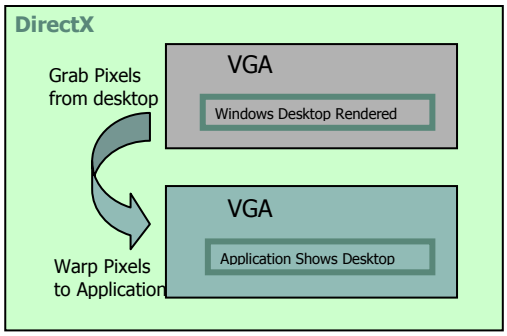


Fig. 4 Postwarp process

2.1.3 Alignment

Before warping process, both pre-warping and post-warping, we must find the parameter for warping. We have developed software for automatic extraction of warping parameter. Sobel first derivative operators are used to take the derivatives x and y of an image, after which a small region of interest is defined to detect corners in. A 2x2 matrix of the sums of the derivatives x and y is subsequently created as follows:

$$C = \begin{bmatrix} \sum D_x^2 & \sum D_x D_y \\ \sum D_x D_y & \sum D_y^2 \end{bmatrix} \quad (1)$$

The eigenvalues are found by solving $\det(C - \lambda I) = 0$, where is a column vector of the eigenvalues and I is the identity matrix. For the 2x2 matrix of the equation above, the solutions may be written in a closed form: equation (2)

$$\lambda = \frac{\sum D_x^2 + \sum D_y^2 \pm \sqrt{(\sum D_x^2 + \sum D_y^2)^2 - 4(\sum D_x^2 \sum D_y^2 - (\sum D_x D_y)^2)}}{2}$$

If $\lambda_1, \lambda_2 > t$, where t is some threshold, then a corner is found at that location. This can be very useful for object or shape recognition[5].

After finding the adequate corner points, we calculate homography H for post-warping process[6].

$$H = (h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{33}) \quad (3)$$

$$A = \begin{pmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1X_1 & -y_1X_1 & -X_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -x_1Y_1 & -y_1Y_1 & -Y_1 \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ x_n & y_n & 1 & 0 & 0 & 0 & -x_nX_n & -y_nX_n & -X_n \\ 0 & 0 & 0 & x_n & y_n & 1 & -x_nY_n & -y_nY_n & -Y_n \end{pmatrix} \quad (4)$$

Given corner points, we can make the following A matrix so that $AH = 0$, subject to $|H| = 1$. Finally, we can get the eigenvector (H) of least eigenvalue of $A^T A$.

2.1.4 Blending

Regions of the display surface that are illuminated by multiple projectors appear brighter, making the overlap regions very noticeable to the user. Since we can detect the overlapping area from alignment process, we applied multiple texturing with blending mask and transparent overlay technique to make seamless overlapping area[7].

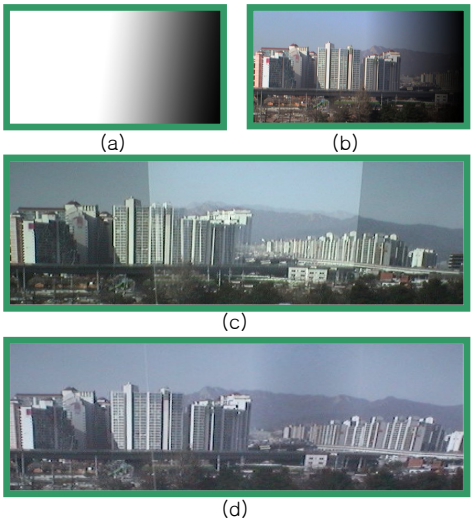


Fig. 5. (a) alpha mask (b) blending with alpha mask (c) without blending (d) final blending image

2.2 Interaction module

In this module, we have integrated with video module which provides multiple camera inputs. Using those inputs, we can detect the users and track the position of users dynamically. And also, extract the information of dynamic hand gesture for interaction with the synthetic object.

2.2.1 Vision based tracking

We use a cascade of boosted classifiers for rapid object detection[8]. In this method, we use a set of features which are reminiscent of Haar Basis functions. Any of these Haar-like features can be computed at any scale or location in constant time using the integral image

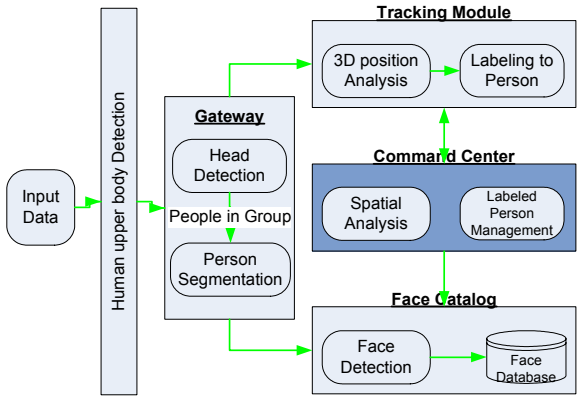


Fig. 6. Diagram for vision based human tracking

representation for images. After detecting the center position of human in the input images, we analyze the 3D position of users and labeling to the person for multiple users. To analyze the 3D position, we must calculate the perspective projection matrix and perspective distortion matrix. From the epipolar geometry between two or more images, we can compute a projective reconstruction of the scene. To compute the camera projection matrices with respect to a projective basis, we choose four pairs of points as a projective basis, any four points of which are not coplanar, between the two cameras. We select four arbitrary points in the reference frame, noted by m_i. For each point m_i, its corresponding epipolar line is given by $l_i^{'} = Fm_i$. We can now choose an arbitrary point on $l_i^{'}$ as $m_i^{'}$, the corresponding point of m_i. Given a pair of points in correspondence: $m = [u, v]^T$ and $m' = [u', v']^T$. Let $\tilde{x} = [x, y, z, t]^T$ be the corresponding 3D point in the space to the projective basis chosen before[9, 10].

$$\begin{aligned} s[u, v, 1]^T &= P[X, Y, Z, 1]^T \\ s'[u', v', 1]^T &= P'[X, Y, Z, 1]^T \end{aligned} \quad (5)$$

We have reconstructed a set of 3D points $\tilde{x}_i = [x_i, y_i, z_i, t_i]^T$ ($i = 1, \cdots, n$) with respect to a projective basis. For each of these points, we know precisely its 3D coordinate in a Euclidean reference frame. Let the set of 3D Euclidean points be $\tilde{X}_i = [x_i, y_i, z_i, t_i]^T$ ($i = 1, \cdots, n$). The projective points $\tilde{x}_i$ are related to the Euclidean points $\tilde{X}_i$ by collineation that we'd like to call the projective distortion matrix D.

$$\lambda_t[x,y,z,t]^T = D[X,Y,Z,1]^T \quad (6)$$

From the equations (5) and (6), we can compute the perspective projection matrix and perspective distortion matrix as follows.

$$\begin{aligned}[u,v,1]^T &= M[X,Y,Z,1]^T \\ [u',v',1]^T &= M'[X,Y,Z,1]^T\end{aligned} \quad (7)$$

where, $[u,v,1],[u',v',1]$ are the detected position of image plane, M and M' are the matrices that are composed of scaling factors, perspective projection matrix and perspective distortion matrix., $[X,Y,Z]$ is a real 3D position what we want to find. We, finally, can find the 3D position with upper equations by least square method.

2.2.2 Volume based motion capture

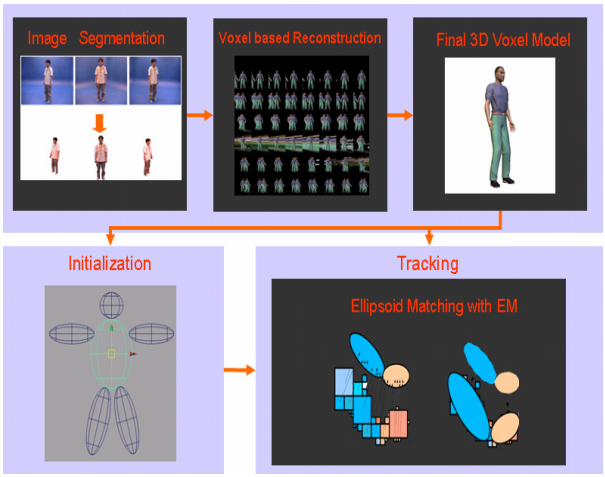


Fig. 6. Flow of volume based motion capturing

For user's interaction with synthetic object, we capture the 3D motion information using volume based method. We use ellipsoid matching with EM method[11] to capture the motion. To do this, we must initialize the initial pose of human. We calculate the 3D location of the voxel set's center of gravity and set the bounding box of torso. And then we estimate the left, right arm and head position with respect to skin color and location information. We, finally, calculate the left and right leg with respect to location information. We can estimate the mean color and the corresponding covariance matrix for each body part which is used for EM process. The process is used to match ellipsoids to groups of voxels. In the Expectation process, we compute the distance to every ellipsoid using Mahalanobis distance for each voxel v and then, assign the voxel to the nearest ellipsoid. Eq. (8) :

$$dst(v,E_j) = (P_v - m_j)K^{-1}(P_v - m_j)^T + (C_v - c_j)Kc^{-1}(C_v - c_j)^T$$

where, P_v, C_v are position and color of v, m_j, c_j are mean position and color of jth ellipsoid E_j. K, Kc are the corresponding covariance matrix. In the Maximization process, we estimate the new means and covariances of each ellipsoid using the set of voxels assigned to it.

2.3 Rendering Module

In this module, we make the virtual environment as a model of the real world and live avatar for remote users to interact the same virtual environment.

2.3.1 Panoramic rendering using active network camera

For virtual environment creation, we adapt image based rendering technique, that is, panoramic rendering. To make panoramic image with remote real world, we utilize the active network camera. Using this camera, we can save several camera resources and also spare the network bandwidth. PTZ functionality expands the field of view of camera and can compensate for the case of using several cameras. We assume that the remote background does not change very fast. We make our camera configure dynamically by preset functionality and therefore, can overcome the limitation of using only one camera based on the assumption.

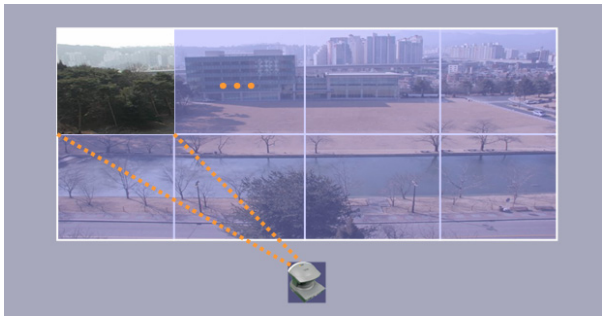


Fig. 7. Panoramic image creation with network camera

Our active camera configures 4x2 presets (each preset's resolution is VGA) so that the image resolution of final composition is over full High Definition resolution. Due to the preset functionality, we can have pre-calculated homographies between presets of active camera. To construct panoramic image mosaics from sequences of images, we associate a pre-calculated homographies with each input image. In order to reduce accumulated registration errors, we apply global alignment to the whole sequence of images, which results in an optimally registered image mosaic. We, also, develop a local alignment technique which warps each image based on the results of pairwise local image registrations. By combining both global and local alignment, we significantly improve the quality of our image mosaics[12,13]. Because of streamed image sequence by active network camera, we can update dynamically the remote real world scene without modifying the pre-calculated homographies between preset images.

2.3.2 Video avatar creation

To visualize the remote users in virtual environment of remote real world, we make them as live avatar[14]. In order to make a live video avatar, the composition engine in our rendering module captures the image of user from network camera. Composition engine extracts the image of remote users from background and uses it as live video avatar. The video avatar is overlaid on virtual environment which is made by panoramic rendering of remote world. The user can control the position of the video avatar by walking around the floor, and can be tracked in real-time by interaction module.

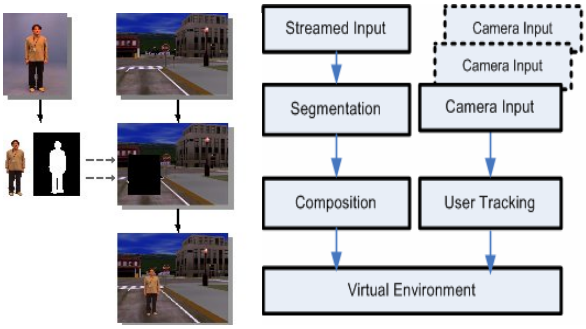
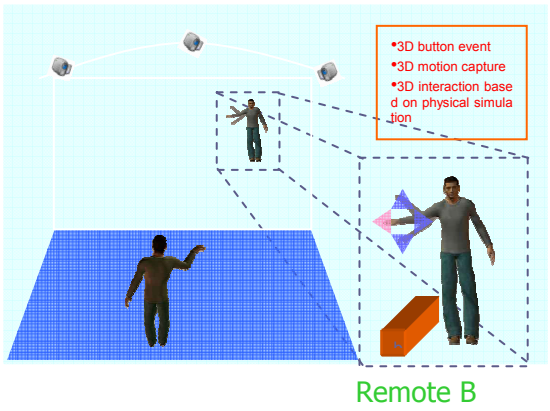


Fig. 8. Video avatar creation process

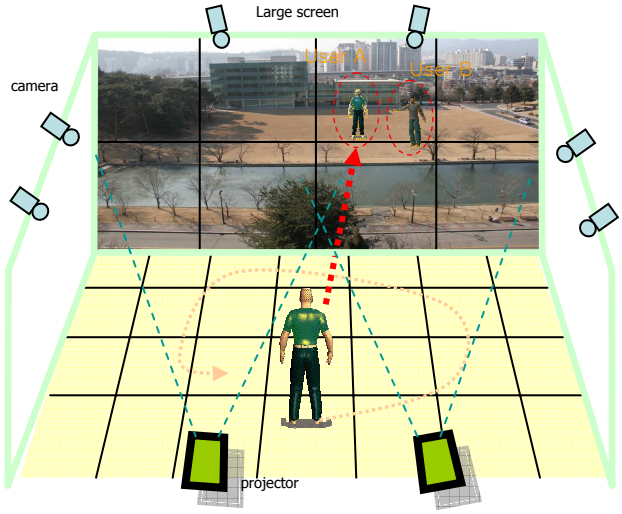
3. Augmented telexistence in smart space

Fig. 9 shows conceptually how to interact between remote users in virtual environment. Fig. 9 (a) shows the remote users can interact with the synthetic object, such as 3D button to touch using the result from interaction module for our scheme. Fig 9(b) shows the example that the remote users can locate the same virtual space as live video avatar. In this case, user A is in smart space and user B is located remote place. Video avatar of user B is transferred and composed by network streaming and rendering module. Before starting the teleconference between remote users, initialization process must be done. In initialize process, the initial position of remote users can be calibrated to fit different space into the virtual space, so that users can interact without misalignment.

In interaction module, we can detect 3D motion information of users using multiple cameras so that we make users interact with the synthetic object based on physical simulation. Physical based simulation can help users to interact very naturally and can give enhanced immersiveness. And also, due to the characteristic of live video avatar, users can watch themselves directly in virtual environment on the large screen. This can make users feel very intimately and interact with remote others on the same virtual space intuitively. Users can change their meeting place dynamically anywhere as long as network camera is connected and the display module can adjust the changed virtual environment for users. In this way, we can make more natural and interactive environment for telecommunication between remote places.



(a) Remote user's interaction



(b) Integrated environment

Fig. 9. Telexistence in smart space

4. Conclusions

In this paper, we describe our environment in detail for the augmented telexistence using natural interaction, seamless large scale display and realistic virtual environment of real remote world. To do this, we have to integrate our sub modules, such as display module for large scale display system, interaction module for physically based dynamic interaction, video module for streaming and rendering module for augmented composition in our smart space. Those systems are well-integrated and allow us to feel augmented telexistence functionality.

Based on our proposed environment, we can create new possibilities for immersive and interactive communication across the space. For the complete solution for telexistence, several things will have to be

done. For example, we are implementing smart floor to detect user's position more precisely and also, have to integrate the dynamic multichannel sound module and haptic module to give a more realistic response to the user.

References

1. S. Tachi, Telecommunication, Teleimmersion and Telexistence, IOSpress, June 2003.
2. S.Tachi, "Toward the Telexistence Next Generation," International Conference on Artificial Reality and Telexistence(ICAT), Invited talks, Dec, 2001.
3. T. Richardson, Q. Stafford-Fraser, K. R. Wood and A.
Hopper, "Virtual Network Computing," IEEE Internet Computing, Vol. 2, No. 1, January/February 1998.
4. http://www.microsoft.com/directX
5. J. Shi and C. Tomasi, "Good Feature to Track," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1994.
6. R. Hartley and A. Zisserman, Multiple View Geometry, Computer Vision, Cambridge University Press 2000.
7. S. C. Ahn, I. J. Kim, H.G. Kim and S. Ha, "Interactive Immersive Display," International Conference on Artificail Reality and Telexistence, TSI Workshop, pp. 63-68, November, 2004.
8. P. Viola and M. Jones, "Rapid object detection using boosted cascade of simple features," IEEE Conference on Computer Vision and Pattern Recognition, 2001.
9. S.M.Yoon, I.J.Kim, S.C.Ahn and H.G.Kim, "Stereo vision based 3D input device," IEEE Conference on Acoustics, Speech and Signal Processing, Vol. 2, pp. 2129-2132, 2002.
10. Z.Zhang, O.D.Faugeras, and R.Deriche, "An Effective Technique for Calibrating a Binocular Stereo Through Projective Reconstruction Using Both a Calibration Object and the Environment," VIDERE1:1, pp58-68, 1997.
11. Dempster, A., Laird, N., and Rubin, D. "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Series B, 39 (1):1–38, 1977.
12. H. Y. Shum and R. Szeliski, "Construction of panoramic mosaics with global and local alignment," International Journal of Computer Vision, 36(2):101-130, Feb. 2002.
13. R. Szeliski, "Image Mosaicing for tele-reality applications," IEEE Workshop on Applications of Computer Vision, pp. 44-53, Dec. 1994.
14. S. C. Ahn, T. S. Lee, I. J. Kim, Y. M. Kown and H. G. Kim, "Computer Vision based Interactive Presentation System," Asian Conference on Computer Vision, pp.486-490, Jan. 2004.