Summary

This paper presents a computer vision based technique for object registration, real-time tracking and image overlay. The capability can be used to superimpose registered medical images such as those from CT or MRI onto a video image of a patient body. Real-time object registration enables an image to be overlaid consistently onto objects even while the objects and cameras viewing it are moving in three dimension.

Reliable real-time object registration is carried out by a sequential process of feature detection in the image, correspondence of those features in the model, and object pose calculation. Image overlay is the projection of models, models to be superimposed, onto image planes with object pose.

Feature detection is executed by computing normalized correlation to reference images at every point in the small search area. The search area is updated every cycle and its repetition in sequential frames realizes feature tracking. The change of appearance of feature points due to view change is compensated by skewing reference images using the object pose information computed in every cycle during tracking.

In feature correspondence, successfully tracked feature points are chosen not only by normalized correlation values, but also by computing variations of geometric invariants from initial values. In the case where feature points do not have simple textures, geometric relationship and constraints between feature points are effective for check of tracking results. “Five Coplanar Points”, which are one of major projective invariants, is used in this paper.

After the feature correspondence, object position and orientation is computed from those feature positions in the image. In the case where we have object-centered coordinates of objects, the object pose can be calculated from a monocular image and the problem is formulated as an inverse problem to solve non-linear relationship between object pose and feature positions in the image. This problem can be solved by recursive methods such as Newton’s method.

Two experimental results to superimpose registered model data are shown. The first example is the tracking of a PC and image overly of the image of an I/O board. The second one is the tracking of a phantom leg with some marks on it and the overlay of a bone model on the view of the leg. These problems are implemented on multiple digital signal processors system with low latency vision hardware. Real-time tracking and image overlay is carried out at frame rate (30 Hz) and overlaid images are kept on the same position in the image during three-dimensional motion of the object such as the PC and the phantom leg, and the camera.

Another type of image overlay system is shown in this paper also. By using properties invariant in view change, overlay of virtual pin attached on a phantom leg in the image is carried out. No model is used in this case. Its portability enables easy application in interactive communication between rural surgeons and experts, which helps delivery of expert care to areas which are geographically or socioeconomically isolated.

has been the principal investigator of several major vision and robotics projects at Carnegie Mellon. He was a founding chair person of CMU's Robotics Ph. D. Program, probably the first of its kind. Dr. Kanade is a Fellow of the IEEE, a Founding Fellow of American Association of Artificial Intelligence, and the founding editor of International Journal of Computer Vision. Dr. Kanade has served for many government, industry, and university advisory or consultant committees, including Aeronautics and Space Engineering Board (ASEB) of National Research Council, NASA's Advanced Technology Advisory Committee (Congressionally mandate committee) and Advisory Board of Canadian Institute for Advanced Research.

ity enables easy application in interactive communication between rural surgeons and experts, which helps delivery of expert care to areas which are geographically or socioeconomically isolated.

References

1. E.R. John, L.S. Prichep, J. Fridman and P. Easton, Neurometrics: computer-assisted differential diagnosis of brain disfuntion, Science Vol. 239, pp.162-169 (1988).
2. L.S. Hibbard, J.S. McGlone, D.W. Davis, R.A. Hawkins, Three-Dimensional Representation and Analysis of Brain Energy Metabolism, Science, Vol. 236, pp.1641-1646 (1987).
3. W.E.L. Grimson, T. Lozano-Perez, W.M. Wells , G.J. Ettinger, S.J. White, R. Kikinis, An Automatic Registration Method for Frameless Stereotaxy, Image Guided Surgery, and Enhanced Reality Visualization, Proc. CVPR'94, pp.430-436, Seattle, WA (1994).
4. D.G. Lowe, Robust Model-Based Motion Tracking Through the Integration of Search and Estimation, Int. J. Computer Vision, Vol. 8, No.2, pp. 113-122 (1992).
5. D.B. Gennery, Visual Tracking of Known Three-Dimensional Objects, Int. J. Computer Vision, Vol. 7, No. 3, pp. 243-270 (1992).
6. A. Rosenfeld and A. Ka, Digital Picture Processing, New York Academic (1982).
7. D.G. Lowe, Fitting Parameterized Three-Dimensional Models to Images, IEEE Trans. Patt. Anal. Mach. Intell. Vol. 13, No.5, pp. 441-450 (1991).
8. S. Yoshimura and T. Kanade, Fast Template Matching Based on the Normalized Correlation by Using Multiresolution Eigenimages, Proc. IROS'94, Munchen, Germany (1994).
9. O. Amidi, Y. Mesaki, T. Kanade, and M. Uenohara, Research on an Autonomous Vision-Guided Helicopter, Proc. RI/SME Fifth World Conf. on Robotics Research, Cambridge, Massachusetts (1994).
10. H. Inoue, T. Tachikawa and Masayuki Inaba, Robot Vision System with a Correlation Chip for Real-Time Tracking, Optical Flow and Depth Map Generation, Proc. IROS'92, Nice, France (1992).
11. I. Weiss, Geometric Invariants and Object Recognition, Int. J. Computer Vision, Vol.10, No.3, pp. 207-231 (1993).
12. J. Munday and A. Zisserman, Introduction-Towards a New Framework for Vision. In Geometric Invariance in Machine Vision, MIT Press, Cambridge, MA (1992).
13. H. F. Durrant-Whyte, Uncertain Geometry in Robotics, IEEE J. Robotics and Automation, Vol.4, No.1, pp.23-31 (1988).

About the Author---Michihiro Uenohara received the B.S. degree in Electrical Engineering from the University of Tokyo, Tokyo, Japan, in 1985. He joined Toshiba R&D Center, Kawasaki, Japan in 1985. He is currently a visiting research scientist at Carnegie Mellon University, Pittsburgh, Pennsylvania. His research areas of interest are computer vision and robotics. He is a member of the IEEE.

About the Author---Takeo Kanade received his Doctoral degree in Electrical Engineering from Kyoto University, Japan, in 1974. Currently, he is U. A. Helen Whitaker Professor of Computer Science and Director of the Robotics Institute at Carnegie Mellon University. Dr. Kanade has made technical contributions in multiple areas of robotics: vision, manipulators, autonomous mobile robots, and sensors. He has written more than 150 technical papers and reports in these areas. He

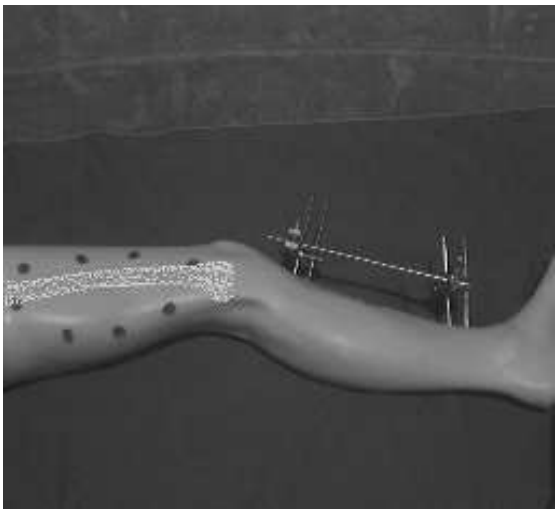

Fig.6 Overlay of a bone on a leg.

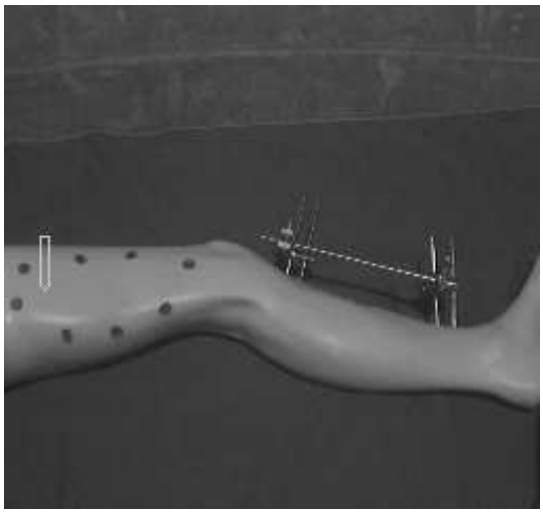

Fig.7 Overlay of a virtual pin on a leg.

Pin Overlay without Models

Image overlay of a virtual pin onto a phantom leg is tested (Fig.7). An overlaid image of the pin remains fixed onto the image of the leg over some motion of the leg. The tip of the pin is supposed to be attached on the leg. Four marks around the pin tip are kept tracking and the position of the pin tip is computed directly from these 2D mark positions. No model is used in this case. Only the pin tip position in the initial image is given.

The pin tip and the four marks around it are almost coplanar in this case. Five coplanar invariants computed by eq. (2) in the initial image with given pin tip position and four mark positions stay invariant over leg motion and camera motion. There are two invariants I_1, I_2 so that x, y coordinates of the pin tip in each frame can be calculated with four mark positions in each frame and the values of I_1, I_2 in the initial image by solving I_1, I_2 in x, y coordinates of the pin tip.

The pin is displayed vertically in the image every time and only the tip position is computed from tracking results. The initial pin tip position is given in advance in this experiment. In the case of interactive video, when experts touch the screen to indicate the specific position of patients' bodies, the touched position is transferred to the remote site as the given pin tip position and the virtual pin is superimposed on the image of patients. Surgeons can recognize the place on patients where experts point to even after some motion of the patients' bodies.

Conclusions

This paper has presented an image overlay system that uses computer-vision based object registration and tracking. Reliable real-time object registration is realized by the tracking of many features on the object and the correspondence of features using geometric invariants in combination with normalized correlation value. Real-time tracking of objects and overlaying image at frame rate (30 Hz) is achieved by the multiple DSP system with low latency vision hardware. It produces natural overlaid images without any delay. They are kept on the same position of the patient in the image as if it is physically attached during three-dimensional motion of the patient and a camera. Computer vision realizes real-time registration of pre-operative model data such as 3D bone and organ model derived from CT, MRI to intra-operative surgical data.

Another type of image overlay system is shown in this paper also. By using properties invariant in view change, overlay of virtual pin attached on a phantom leg in the image is carried out. No model is used in this case. Its portabil-

Fig.5a Overlay of the board (No.1).

Fig.5b Overlay of the board (No.2).

position in the last frame where the normalized correlation score is the highest. However, in the case where the feature point is missed in the last frame, the projected feature position in the image, computed with the pose of the PC in the last frame, is used to define the center position of the search area. This allows for recovery of tracking.

Eight tracking results are checked by calculating five coplanar invariants. All the eight features are located on the same plane. The number of combinations of eight points taken five points at a time is 56. Therefore, 56 invariants are calculated and differences from the initial values divided by standard deviations are computed. The best five points which have minimum change is selected, and the pose of the PC is computed with these five points by Newton's method. Feature points which have a peak under 0.7 are rejected before the computation of invariants. Peaks when the feature points are missed may exceed 0.7. Normalized correlation peak score is not enough for selection of successfully tracked features when the texture of feature points and background are not so simple: A check by geometric invariants, combined with normalized correlation peak score, makes the tracking much more robust. The system can track and superimpose the I/O board in three-dimensional translation and rotation of the camera and the PC. It is capable to keep tracking and superimposing even when up to three feature points are occluded by other objects, and human hands. The tracking and overlay of the board is executed at frame rate (30 Hz). Three C40s are used in parallel: Two C40s are used for tracking of eight feature points and one is used for check of tracking results and pose calculation and image overlay.

Overlay of an Image of a Bone onto a Leg

The Image overlay of a bone onto a phantom leg is carried out by putting eight fiducial marks on a leg (Fig. 6). The bone surface model is derived from CT data. The fiducial marks should be put on the leg keeping consistency of two coordinates used for the bone model and the marks model which includes the local coordinate positions of each mark. In this case, no feature correspondence by geometric invariant is carried out and successfully tracked features are selected only by normalized correlation value. Since there are no complex features around marks and normalized correlation gives us reliable enough information. As in the PC case, the overlaid image of the bone remains attached to the leg in three-dimensional camera motion and leg motion. The system keeps tracking and superimposing bone model in some occlusion also by selecting non-occluded marks and computing object pose by those mark positions in the image.

high speed data link. Image data are transferred through this high speed data link into C40 processor communication ports, and then transferred to the local memory of the processor and other processors' communication ports by DMA. Multiple C40s and this low latency vision hardware have enabled us to achieve real-time visual tracking of objects and image overlay at frame rate (30 Hz).

Overlay of an image on a PC

Visual tracking of objects without attaching specific marks is tested on a desktop PC. The PC is uncovered and located in front of the camera, about 1.4 meters away. At the beginning of the operation, the system displays a wire frame display of the PC on the monitor. The system requires a user to move the camera so that the PC and the wire frame are aligned (Fig.3). When the PC is roughly aligned to the wire frame, the system recognizes it, "latches" onto it, and starts tracking it. Initial recognition is executed by template matching of three regions on the PC. The measure of similarity in the matching is normalized correlation. Eleven images of the PC in different illumination are precaptured and three 32 x 32 regions are extracted from each image as reference images. Initial recognition succeeds when the two conditions are satisfied: (1) The first one is that the best peak value of normalized correlation is over the thresholds at all three regions. The thresholds are set to be 0.80. (2) The second one is that the peak is not located on the edge of the search area. A peak on the edge of the search area may mean that there is a real peak outside of the search area.

The use of reference images in various illumination makes the recognition robust for light condition change, yet requires more computation. Reference images are slightly different from each other, but basically similar and highly correlated. They are compressed into an average image and four eigenvectors by Karhunen-Loeve expansion and are downloaded at the execution. This compression makes template matching nearly three times faster.

When the two criteria described above are satisfied, the pose of the PC is calculated from three feature positions in the image by Newton's method. Feature points for tracking are projected onto the image with the computed pose and small windows of size 16 x 12 around feature points are extracted from the image and are used as reference images in the tracking after that (Fig.4). The number of feature points in the tracking phase is eight. During tracking (Fig.5), eight feature points are tracked by computing normalized correlation to reference images at every point in the small search area and the points with the highest scores are selected while deforming reference images with object pose. The search area in the tracking phase is 14 x 14, and the center position of the search area is usually set to the feature

Fig.3 Wire-frame overlay at initial recognition.

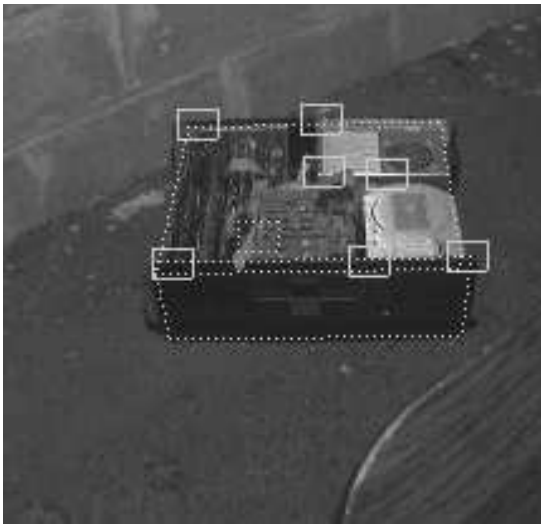

Fig.4 Reference images during tracking.

where Λ_I is the covariance matrix of invariants and Λ_p is the covariance matrix of feature positions in the image. $\boldsymbol{J}$ is the Jacobian matrix $\partial I/\partial \boldsymbol{x}$ and $\boldsymbol{x}$ is the feature position vector whose elements are x-y coordinates of five feature points. We assume that there is no correlation between observation error of each point so that Λ_p becomes diagonal matrix and Λ_I becomes scalar, which is expected variance σ_I^2 of invariants.

Threshold for each invariant is set to the standard deviation of invariants σ_I multiplied by some constant c. Five-coplanar invariants are computed for all combinations of five feature points out of all feature points. They are compared with their thresholds and discarded if variations from initial values are over their thresholds. When there are more than one invariants under thresholds, five feature points which have the minimum variation of invariant divided by the threshold are selected. Object pose is calculated with these five feature positions in the image.

Pose Calculation

Projection from 3D to 2D is a nonlinear operation. It is hard to solve directly the inverse problem from 2D feature point positions in the image to 3D object pose. Fortunately, however, it is a smooth and well-behaved transformation. Therefore, Newton's method can be applied. While it requires an appropriate initial estimate, we can get a good estimate of the pose parameters from the results of the last frame during tracking.

Newton's method computes a vector of corrections $\boldsymbol{x}$ to be added to the current estimate of pose parameters $\boldsymbol{p}$ on each iteration. Given a vector of error measurements $\boldsymbol{e}$, whose elements are x, y coordinates of errors between measured feature positions and projected feature positions computed with the current estimate of the pose. $\boldsymbol{x}$ is solved in the following matrix equation:

$$\boldsymbol{J}\boldsymbol{x} = \boldsymbol{e} \tag{4}$$

where $\boldsymbol{J}$ is the Jacobian matrix $J_{ij} = \partial x_{f_i}/\partial p_j$ and elements of $\boldsymbol{x}_f$ are x, y coordinates of feature points in the image.

If the problem is locally linear, the error will be reduced to zero after adding the corrections. When five feature points are used for pose calculation, there are more measurements than parameters and this system of equations becomes overdetermined. we can get an x that minimizes the norm of the residual by the normal equations:

$$\boldsymbol{J}^t\boldsymbol{J}\boldsymbol{x} = \boldsymbol{J}^t\boldsymbol{e} \tag{5}$$

As long as there are much more measurements than parameters, Newton's method will usually converge in a stable manner from a wide range of starting positions. Numerical stabilization methods, which add prior constraints such as zero corrections to the current pose parameters can be applied also[4, 7].

Experimental System for Real-Time Image Overlay

The techniques for object tracking and image overlay presented above have been implemented. We will present three examples. The first example is the tracking of a PC and image overlay of the image of an I/O board. The second one is the tracking of a phantom leg with some marks on it and the overlay of a bone model on the view of the leg. And the last one is the overlay of a virtual pin onto a leg model.

The systems are implemented on multiple TMS320C40 (C40), Texas Instruments digital signal processors. C40 is effective for parallel processing and real-time embedded applications. Its advantages are high speed communication ports each capable of transferring data at 20 MB/s, programmable DMA in addition to high speed floating-point calculation. We use low latency vision hardware developed at CMU[9] which has a digitizer, convolution hardware and

where $\boldsymbol{x}_{ci}$=[x_{ci}, y_{ci}, z_{ci}] are the camera-centered coordinates, $\boldsymbol{x}_{li}$ are the object coordinates stored in the model, R is a 3 x 3 rotation matrix defining object orientation and T is a translation vector defining object position, and f is the focal length, respectively. The z axis in the camera-centered coordinate coincides with optical axis of the camera. R and T can be calculated from object pose so that $\boldsymbol{x}_{fi}$ can be represented as a function of $\boldsymbol{x}_{li}$ as $\boldsymbol{x}_{f\,i} = \boldsymbol{g}\ (\boldsymbol{x}_{li})$. $\boldsymbol{x}_{f\,i}$ is easily computed from feature positions. The function can be rewritten as $\boldsymbol{h}(x_{li}) = 0$, which is a linear equation of dimension two and $\boldsymbol{x}_{li}$ is the three dimensional solution vector. Surface patch equations $a\,\boldsymbol{x}_{li} = 0$ are given as part of models also so that $\boldsymbol{x}_{li}$ can be solved from these equations. $\boldsymbol{x}_{li}$ can then be easily projected onto the initial image by the pose of the object at the initial recognition.

Feature Correspondence

While tracking, we can reasonably assume that the correspondences of feature points in the last frame is mostly maintained in the current frame. Some features, however, may be missed or mismatched during tracking. So we need to check and select only successfully tracked feature points. Criterion is used in such a selection process.

The peak value of normalized correlation is one of candidates. It indicates the similarity of current images and reference images. The distribution of correlation values over the search area contains useful information[10]. If the distribution has a single sharp peak, the feature is likely found successfully, while homogeneous distribution indicates unreliable matching. These kinds of checks are adequate for tracking of good simple features, such as special marks put on objects for tracking.

Sometimes it is not enough to evaluate the degree of matching at the image level. The appearance of features varies, for example, by illumination change. Normalized correlation yields many peaks and some of them have larger scores than the correct one when the appearance of features changes from the original. Geometric relationship and constraints between feature points should be used to cope with these difficulties.

Geometric invariants[11, 12], popular in object recognition as useful descriptor of objects, are properties in the image that stay invariant under some transformation. One major invariant is “Five Coplanar Points”. The cross-ratios of four areas of triangles S_{ijk} as below are invariant under projective transformation.

$$I_1 = \frac{S_{423}S_{125}}{S_{124}S_{523}} \qquad I_2 = \frac{S_{143}S_{125}}{S_{124}S_{153}} \tag{2}$$

S_{ijk} is the area of a triangle with three points, i, j, and k. These values remain the same value over view changes. If the geometric invariant values computed from tracking results change, it indicates that some of feature points are misrecognized. “Five Coplanar Points” are good invariants for objects with many planar surfaces. Calculation cost to compute invariants is relatively small also. Reliable tracking is accomplished by the combination of tracking of many feature points, check by invariants, selection of successfully tracked feature points and pose calculation by those good feature points.

One important point we must concern with is sensitivity of geometric invariants. Some observation errors are included in tracking feature positions (typically 0.5 pixel), which cause invariants computed from them to vary. The sensitivity of invariants is also dependent on configuration of feature points. It makes difficult to set constant thresholds to judge whether or not invariants are violated and thus feature points are successfully tracked. Instead of constant thresholds, thresholds need to be adjusted by the standard deviation of each invariant. Assume that observation errors of each feature position has a Gaussian distribution. Invariants computed by (2) then have the distribution with covariance matrix as below up to second order[13]

$$\Lambda_I = J\Lambda_p J^t \tag{3}$$

tracking phase. The minimum number of feature points required in initial recognition is three. This is the minimum number necessary for calculating the 3D pose of objects. It is assumed that there is no occlusion at the beginning of the execution. The initial pose calculation is by the same principle as in the tracking phase, which is described later.

Tracking of Features

Feature points that are easy to track are selected before execution, and each position of the feature points in object coordinates is given as part of the object models. When the object is recognized and the pose is calculated at the initial recognition phase, feature points are projected onto the image plane with the computed pose and the small regions around the feature points are extracted as reference images for the subsequent visual tracking.

For visual tracking, normalized correlation to reference images is computed at every point in the small search areas. The positions with the best normalized correlation scores are selected as the positions of feature points in the image. In the next frame, search occurs in the small area around the selected point, and this process repeats.

The appearance of feature points varies during tracking due to view change and by illumination change, and other reasons. The values of normalized correlation is insensitive to the intensity changes, which makes it robust against illumination changes.

The change of appearance of feature points due to view change is compensated by skewing reference images using the object pose information computed in every cycle during tracking. Reference images are small square windows of N by N pixels around feature points. In the perspective transformation, the region around a point of size N by N may be deformed to different shapes. However, straight lines are projected to straight lines and intersections are preserved in any view change. Skewed reference images are generated by computing four corner positions in the initial image which correspond to the four corners of the small square window around the feature point in the current image, and packing pixels in the region surrounded by the four straight lines connecting the four corners in the initial image into a square area of N by N pixels (Fig.2). The four corner points in the initial image are computed as below. Surface patches around the feature points are approximated as planar surfaces and those equations are given as part of object models.

Four corner point positions $\boldsymbol{x}_{f_i}$ in the image plane are computed by

$$\boldsymbol{x}_{f_i} = (f/z_{c_i}) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \boldsymbol{x}_{c_i}$$
$$\boldsymbol{x}_{c_i} = R\boldsymbol{x}_{l_i} + T \tag{1}$$

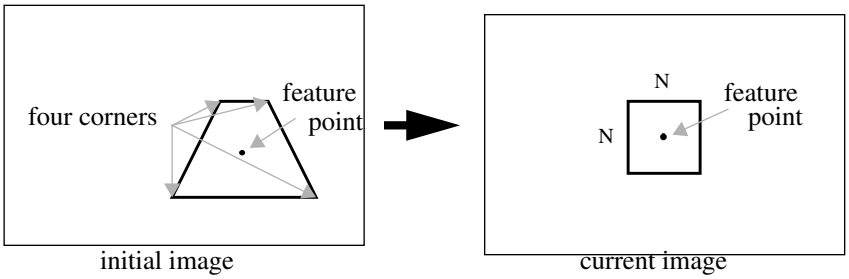

Fig.2 Change of appearance of feature points.

When the 3D pose of an object has been computed, the last remaining step is to generate an image of prestored data, such as pre-operative 3D bone model derived from CT, appropriately projected onto the current image plane, and add it to a raw camera image. This way, computer vision enables real-time registration of pre-operative to intra-operative surgical data.

There may be another type of image overlay technique that does not require object models. For the interactive video, we just want the system to keep tracking the specific point or contour and to superimpose it on the same place on a patient's body in the image. In this case, we don't need the model of objects to superimpose. Also, we don't need the object model for registration if we can compute the position of point or contour in the current image just by keeping track of some neighboring features and computing the position of the point or contour directly from those feature positions.

There are certain image properties which are invariant to changes of viewpoints and help us compute the position of points without registration. The most representative invariants are cross ratio of four points and cross ratio of areas of five coplanar points. They are invariant in any view change, and enable us to calculate the position of one of the points from the other 2D feature positions in the current image with the value of invariants in the initial image.

Vision-based real-time object registration which realizes image overlay of pre-operative model data is described first. Then, direct computation of image overlay without models is discussed. Real-time object registration is divided into the initial recognition and tracking. The main difference between them is whether we have a good estimate of the pose of objects to start with. Initial recognition is described in the next section, and the processes for object registration are described thereafter. Some developments of image overlay systems are shown at the end of the paper.

Initial Recognition of the Object

To start visual tracking, the object must be localized first in the initial image. The major difficulty in initial recognition is that information about object pose and location in the previous image cannot be used as a starting pose estimate and reference image. In general, the appearance of objects changes in various ways depending on pose and illumination change, which causes also problems.

For human interface systems, however, it is a reasonable assumption to think that users can locate objects at the start of the execution of the system. The system, for example, can show the desired position and orientation of the object by superimposing some images onto the raw camera image. Users can set the pose of the object by moving the object or by moving the camera.

When the pose of an object is initially set by user as described above, feature points appear at the predefined positions in the image so that initial recognition is carried out by a search of the feature points within limited areas around those predefined points in the image. Reference images around feature points are precaptured in various illumination while objects are set to be the predefined pose. Normalized correlation is computed with all these reference images at each point in the search area. The point with the highest score is chosen, and it is recognized as a feature point when the highest score is over a threshold.

The robustness of the initial recognition against illumination changes improves by using multiple reference images in different illuminations. It requires us, however, more computation. We reduce computational complexity by approximating reference images as linear combinations of some major eigenvectors. Karhunen-Loeve expansion gives us optimal major eigenvectors in the sense that the norm of approximation errors become minimum[8, 9]. The number of eigenvectors necessary for approximation is much less than the number of reference images because of the high correlation between reference images. This compression greatly reduces the computation cost of normalized correlation when the number of pixels in each reference image is much larger than the number of reference images, which is generally the case.

When all the feature points are successfully found, the pose of the object is calculated and the system goes to the

Object Registration for Image Overlay

Computer-vision based object registration is a matching-process between models and images. In the case of intensity images, models are in general sets of features and those relationship. Features are local two-dimensional patterns which are characteristic and easy to distinguish. Those relationships stored in the model are geometric in addition to symbolic. Object registration is therefore a process of feature detection in the image, correspondence of those features to features in the model and object pose calculation as shown in Fig.1[4][5]. Image overlay is the projection of geometric models, models to be superimposed, onto image planes using object pose.

Feature detection is carried out at first stage of the process. In the case where we have given pattern of images, the simplest and attractive approach is template matching, which finds a pattern in the image that is similar to a reference image[6]. Similar features to the reference images are searched in the image by using correlation or normalized correlation as the measurement of similarity. Reference images are precaptured before execution in some cases, or they are extracted in the last frame of images and updated in every cycle in the other cases. In general, increasing the number of features to be detected improves the system's reliability. Due to illumination, occlusion and pose changes, some features may be missed and some wrong features may also be detected.

Once some features are detected in the image, feature correspondence is carried out: i.e., associating a detected feature with the equivalent feature in model. It is executed by generating a hypothesis which features in the image correspond to which features in the model, and then doing verification based on geometric relationship among them. This task is much easier in object tracking than in general object recognition, because feature correspondences in the last frame can be a very good initial estimate for the next frame in tracking. It is, however, necessary to verify successfully tracked features in the case of tracking, as well, because some features may be missed by illumination changes or for other reasons.

After the feature correspondence, object position and orientation is computed. In the case where we have object-centered coordinates in the models, the object pose can be calculated from a monocular image and the problem is formulated as an inverse problem to solve the non-linear relationship between object pose and feature positions in the image. Non-linearity comes from perspective projection. This problem can be solved by recursive methods[7]. If the visual tracking is at a relatively high rate, the change of object pose between cycles is not large and the algorithm tends to converge within a small number of iterations. The minimum number of correspondences necessary for 3D pose calculation is in general three.

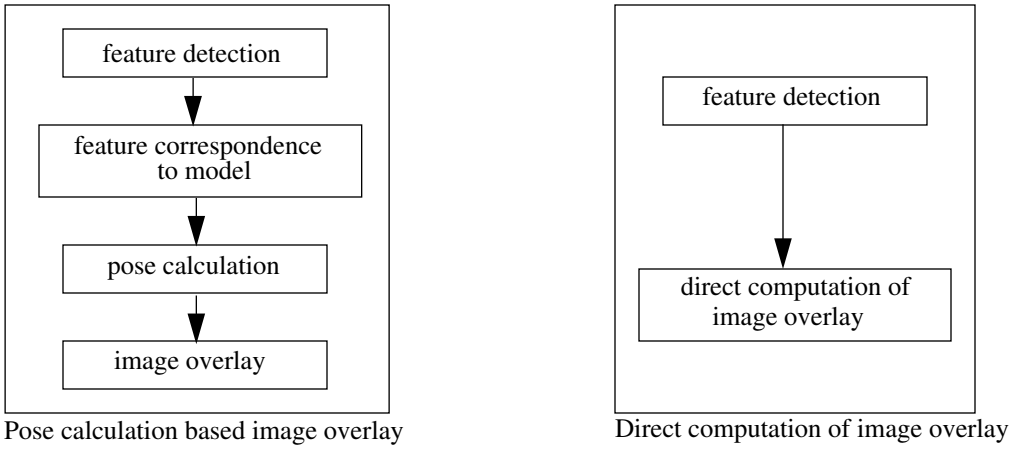

Fig.1 Pose calculation based image overlay and direct computation of image overlay.

Vision-Based Object Registration for Real-Time Image Overlay

M. Uenohara and T. Kanade
The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213

Abstract-This paper presents a computer vision based technique for object registration, real-time tracking, and image overlay. The capability can be used to superimpose registered medical images such as those from CT or MRI onto a video image of a patient's body. Real-time object registration enables an image to be overlaid consistently onto objects even while the objects and cameras viewing it are moving. Object registration is composed of feature tracking, feature correspondence, and pose calculation of objects. This technique is based on geometric models of objects, but it can be extended so that some image overlay is possible without a prior model of the object.

image overlay computer vision visual tracking registration
augmented reality

Introduction

Computer vision has found applications in surgery: computer-assisted detection of anatomical and functional lesions, 3D representation and analysis of brain energy metabolism and so on[1][2]. They are, however, limited mostly to off-line presurgical analysis and processing. Due to the significant improvements in computer vision techniques in recent years, real-time and interactive imaging of complex biomedical systems has become great priority within medicine.

One major challenge is to integrate the precise pre-operative information currently found within CT and MRI into intra-operative surgical procedure. Display of correctly registered medical images on a patient provides a new methods of surgical guidance. That can enhance human perception and skills[3]. Most previous methods of registration, however, are either off-line or assume that the patient does not move relative to a room or the bed during the surgery. Real-time computer vision techniques for object registration can realize non-intrusive image overlay without using special positioning devices. The overlaid image can be kept on the same position of the patient in the image as if the overlay is physically attached to the patient. The overlay must remain fixed to the patient even with motions of the patient and a camera.

Interactive video is another challenge. In telemedicine, rural surgeons would send patient records, X-rays and CT scans to an expert surgeon who would use them to plan the operation on a surgical simulator. That expert would send the surgical plan to the remote doctor or medic and guide him through the surgery. The interactive video, which transmits images of a patient to the expert and sends back with some image overlay, enables the expert to guide surgeons as if the expert were across the operating table from him. It could keep showing the surgeon the place on the patient body the expert points to while the patient and the camera are moving in three dimensional.

This paper presents object registration and tracking techniques appropriate for the realization of real-time image overlay. We first discuss computer vision techniques necessary for initial recognition and tracking of objects, and then show real-time image overlay applications that are implemented on multiple digital signal processors with dedicated vision hardware.