



US 20080152200A1

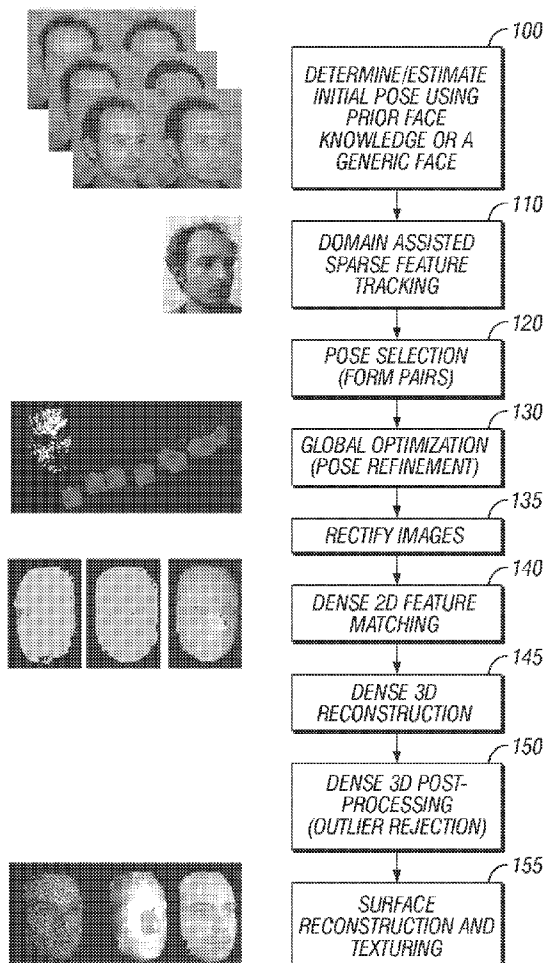
(19) **United States**(12) **Patent Application Publication**
MEDIONI et al.(10) **Pub. No.: US 2008/0152200 A1**(43) **Pub. Date: Jun. 26, 2008**(54) **3D FACE RECONSTRUCTION FROM 2D IMAGES**(75) Inventors: **GERARD MEDIONI**, Los Angeles, CA (US); **Douglas Fideleo**, San Diego, CA (US)Correspondence Address:
KNOBBE MARTENS OLSON & BEAR LLP
2040 MAIN STREET, FOURTEENTH FLOOR
IRVINE, CA 92614(73) Assignee: **Clone Interactive**, Pasadena, CA (US)(21) Appl. No.: **11/828,232**(22) Filed: **Jul. 25, 2007****Related U.S. Application Data**

(63) Continuation of application No. 11/669,099, filed on Jan. 30, 2007.

(60) Provisional application No. 60/764,007, filed on Jan. 31, 2006.

Publication Classification(51) **Int. Cl.**
G06K 9/00 (2006.01)
G06K 9/36 (2006.01)(52) **U.S. Cl.** **382/118; 382/285**(57) **ABSTRACT**

A 3D face reconstruction technique using 2D images, such as photographs of a face, is described. Prior face knowledge or a generic face is used to extract sparse 3D information from the images and to identify image pairs. Bundle adjustment is carried out to determine more accurate 3D camera positions, image pairs are rectified, and dense 3D face information is extracted without using the prior face knowledge. Outliers are removed, e.g., by using tensor voting. A 3D surface is extracted from the dense 3D information and surface detail is extracted from the images.



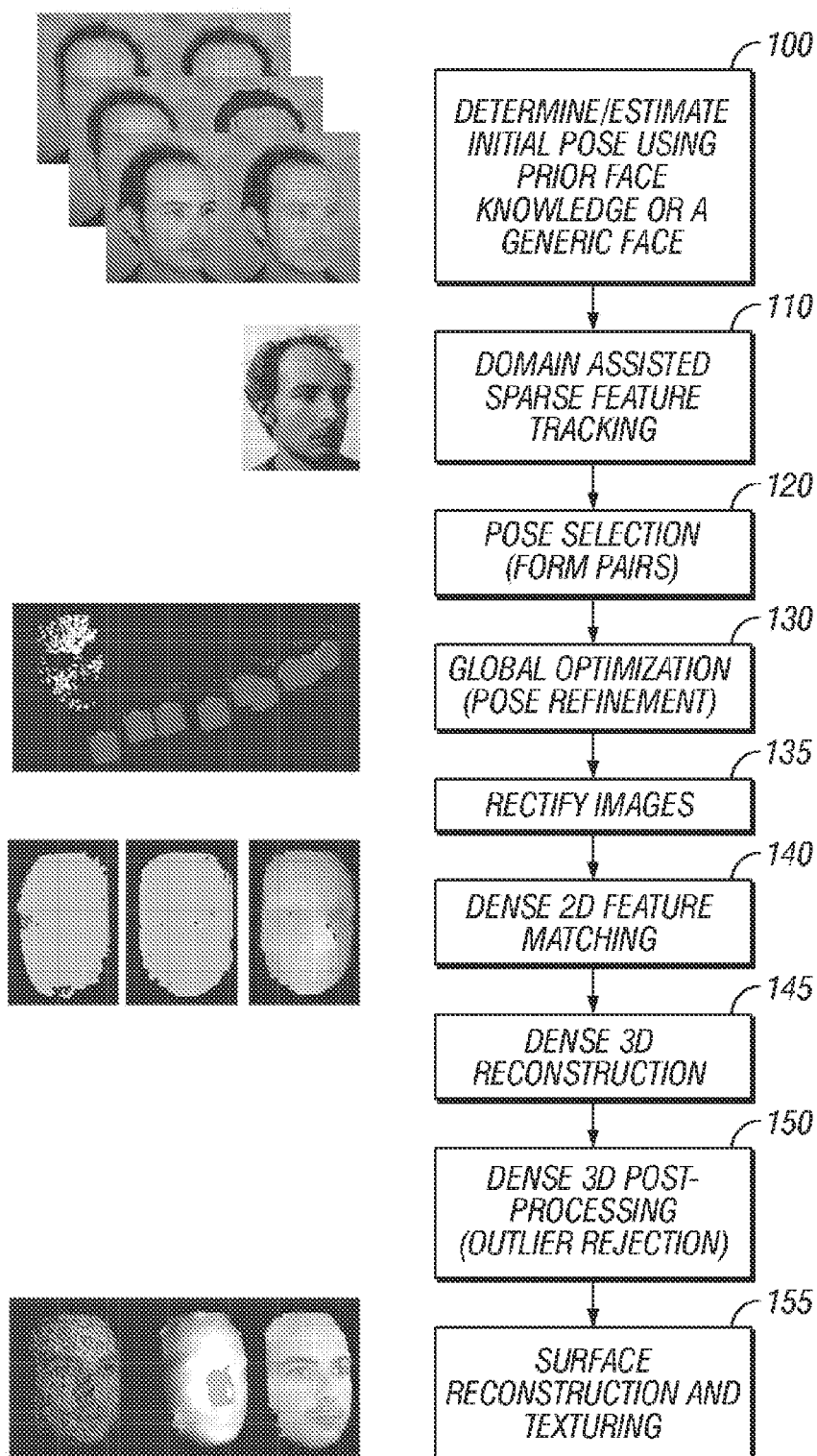


FIG. 1

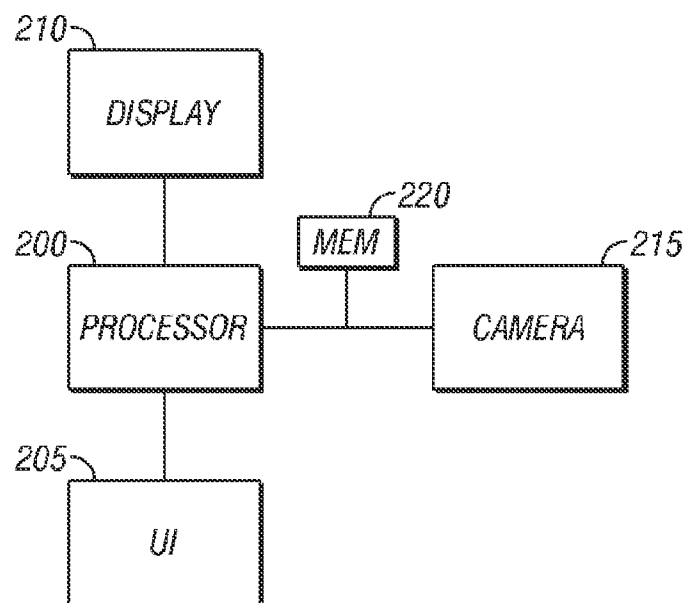


FIG. 2

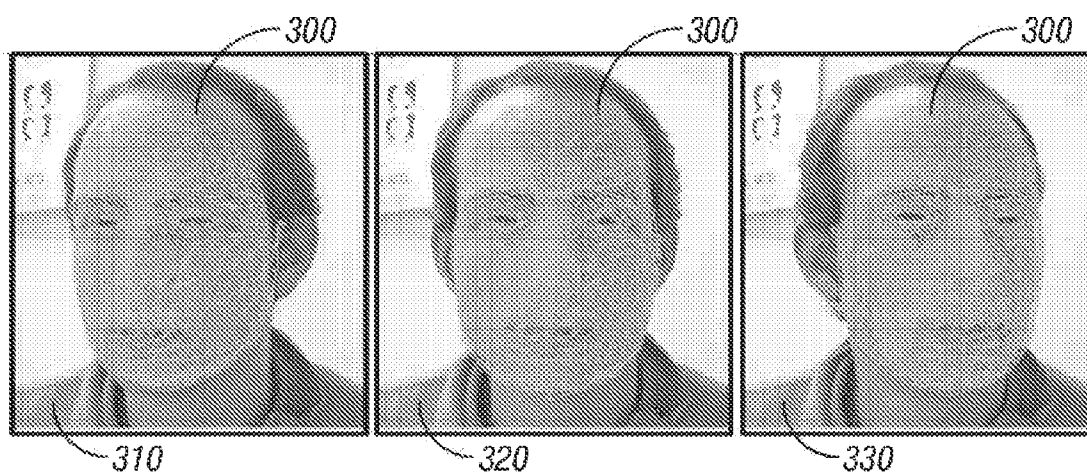


FIG. 3

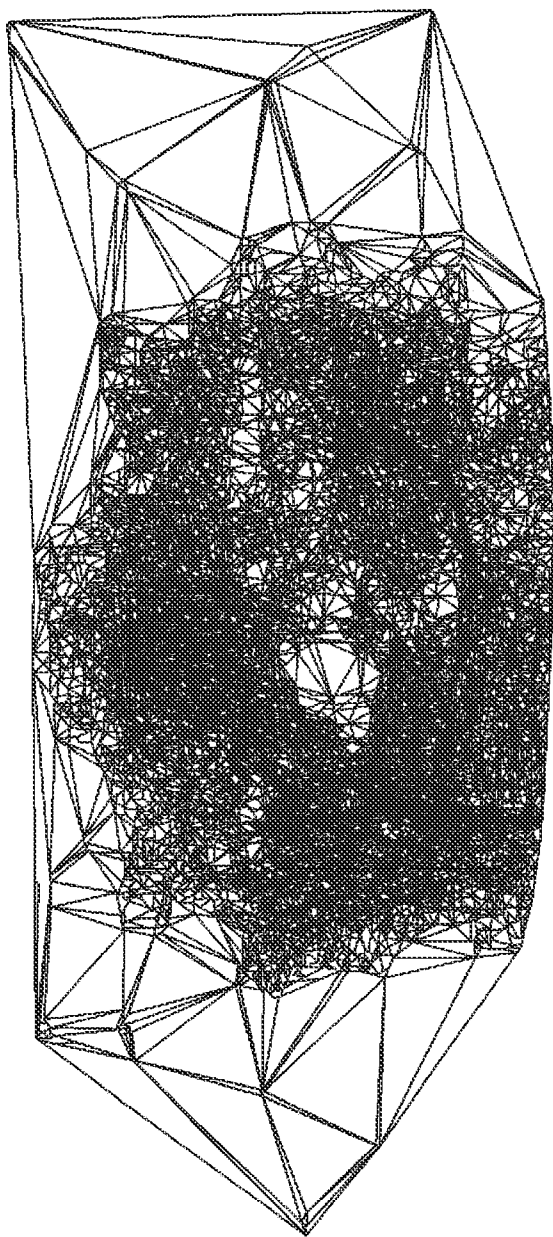


FIG. 4B

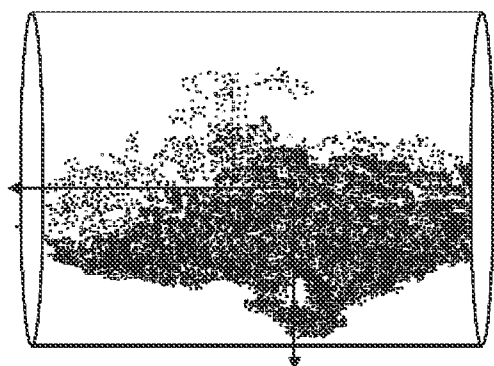


FIG. 4A

3D FACE RECONSTRUCTION FROM 2D IMAGES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to U.S. patent application Ser. No. 11/669,099, filed on Jan. 30, 2007, which claims priority to U.S. Provisional Application 60/764,007, filed Jan. 31, 2006. The disclosure of the prior application is considered part of (and is incorporated by reference in) the disclosure of this application.

FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0002] The U.S. Government may have certain rights in this invention pursuant to Grant No. HMI-582-04-1-2002.

BACKGROUND

[0003] Conventional face reconstruction techniques often use a two dimensional image or images (e.g. digital photographs) of a face to create a three dimensional representation of the face. The representation that is created may be a file, such as an electronic file, indicative of individual characteristics of different faces. The file can then be used, e.g., for facial recognition, animation, or rendering.

[0004] The images, once obtained, are often processed based on prior knowledge or assumptions of what faces usually look like. This knowledge is often called "domain knowledge", a "prior model", or more specifically a "generic face". For example, the prior face knowledge may indicate the presence or likely locations of different kinds of facial features, such as eyes, nose, etc. The prior face knowledge may assume that the face is formed of a linear combination of basis face shapes and appearances, camera parameters, lighting parameters, and other known elements, or elements that are susceptible of estimation. These elements can be combined to estimate the likely appearance of a face. More specifically, the domain knowledge may come in the form of a generic face shape defined by an artist or an average face shape computed from a plurality of known face shapes.

[0005] One common technique for face reconstruction uses prior face knowledge of a generic face, and possibly a set of face metrics or deformation parameters, throughout the reconstruction process. Another common technique attempts to eschew the use of prior face knowledge and instead uses a purely data-driven approach to reconstruct the face. This can be done, for example, using triangulation of two-dimensional points in multiple images from multiple calibrated cameras. Unfortunately, the former approach may provide unrealistic data, due to the use of the generic face throughout the process. The latter approach requires additional hardware infrastructure which is difficult to practically implement at a reasonable cost. A single-camera purely data-driven approach alleviates some of the hardware constraints of multi-view stereo methods, but may itself be unstable due to the lack of constraints at stages of the process.

SUMMARY

[0006] The present application describes techniques for obtaining three-dimensional face information using an assisted technique. According to aspects, prior knowledge of

face structure is used at some points during the processing operation, but other parts during the processing operation are purely data driven.

[0007] Another operation uses a single camera for determination of 3D information from a set of 2D images.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] FIG. 1 shows an overall flowchart of the operation; and

[0009] FIG. 2 shows a general purpose computer which can carry out the flowchart.

[0010] FIG. 3 shows how the three dimensional face tracker is assisted by a generic face.

[0011] FIGS. 4A and 4B show the dense three dimensional features respectively embedded in a cylindrical space and unwrapped and triangulated.

DETAILED DESCRIPTION

[0012] The general structure and techniques, and more specific embodiments which can be used to effect different ways of carrying out the more general goals, are described herein.

[0013] The present application refers to determining three dimensional information about an object, e.g., a face. Although the present embodiment is described with reference to 3D reconstruction and rendering of faces, it should be understood that these same techniques can be used to reconstruct and render multiple views of any object. When used for faces, the three dimensional information produced by the techniques disclosed herein can be used for any face based application, such as animation, recognition, or rendering. The techniques disclosed herein may be more realistic than other techniques that rely more extensively on prior knowledge of generic faces.

[0014] The present inventors recognize that previous systems, which used strong prior knowledge of face appearance to reconstruct a face, in effect quantize the number of basis shapes that are used for forming and rendering the face. A strong prior knowledge or generic face approach is effectively limited by the degrees of freedom that are offered by the imposed prior face knowledge or generic face. Hence, the information and subsequent reconstructions do not capture all the subtle details in the original face.

[0015] This "face space" quantization is caused because the prior knowledge and associated transformations limit the space of all possible faces that can be reconstructed by the system. Generic face or pure prior face knowledge based methods may not have sufficient degrees of freedom to cover the entire face space.

[0016] An embodiment captures subtle face details by disregarding the prior face knowledge or generic face constraints at key points of the process, and instead by relying on the data using a data driven approach to find the details of the face called herein the dense features. The data-driven approach requires large amounts of data to deal effectively with noise, measurement uncertainties, and outliers. However, the present system does not use a purely data-driven approach, but also is assisted by methods that incorporate prior face knowledge or generic faces.

[0017] According to one aspect, large amounts of data can be obtained from a single camera that operates to obtain multiple images. For example, this may use frames of video

which collectively form a moving sequence of images. It may also be obtained from multiple different still images obtained from one or many cameras.

[0018] U.S. Pat. No. 7,103,211 discloses a method for generating 3D face models, that uses no prior face knowledge whatsoever, but rather is completely data driven. The present system uses a system that is mostly data driven, but uses prior face knowledge or a generic face to determine certain parts of the information.

[0019] An embodiment is disclosed with reference to the flowchart of FIG. 1. FIG. 1 also illustrates some exemplary thumbnail images, illustrating the operation. This flowchart can be carried out on any general purpose computer, such as the system shown in FIG. 2. This system includes a processor 200, a user interface such as mouse and keyboard 205, and a display screen 210. The computer can be, for example, an Intel-based processor or any other kind of processor. The computer receives raw or processed image data from one or more cameras 215, e.g. still cameras or video cameras. The processor 200 processes that raw image data according to the descriptions provided herein. As an alternative, the camera information may be stored in a memory 220, e.g. a hard drive, and processed at some later time.

[0020] An embodiment extracts information from a sequence of images, e.g. a video sequence, a sequence of stop motion style images from the video sequence, or simply a number of still images. Unless the subject is standing completely still and the camera does not change positions, the sequence of images will have multiple different views of the subject's head in the set of images.

[0021] At 100, an initial pose estimation is determined. This may use a face tracking algorithm, such as that shown in FIG. 3, to derive an initial head pose estimate, and also to derive a mask which represents the look of the face. This uses prior knowledge of face structure to determine the likely position and pose of the head, location of facial features such as nose, mouth, etc, and the like. FIG. 3 illustrates 3 different poses in three different images 310, 320 and 330. The same face mask 300 is superimposed on each of those images. The face mask 300 represents a generic face, and hence has spots for eyes, nose mouth, etc. In this way, the mask helps estimate the pose of the images.

[0022] The pose estimation technique passes a set of views to a sparse feature tracking module at 110. The views, which are passed to the module, are those which are believed to be good candidates for image pairs from which three dimensional information can be extracted. The sparse feature tracking module 110 produces a set of feature correspondences for each image pair. The two images in a pair are sufficiently close so that these feature correspondences can be obtained.

[0023] Pose selection is carried out at 120, to select those images which properly make a pair that can be used for the determination of 3D information. These pairs should be close in pose and have similar lighting characteristics.

[0024] Global optimization is performed over the entire set of feature points at 130. This is used to refine the camera position estimate and compute the three-dimensional structure of the sparse two dimensional features.

[0025] The refined camera positions are used to rectify pairs of images at 135, thereby constraining the search space for corresponding feature points to a horizontal scan line in the paired images.

[0026] At 140, dense feature matching is performed across the pairs. This finds additional features beyond the sparse

detection that was carried out at 110. These correspondences are determined by triangulation using optimized camera poses to form a dense 3-D point cloud or disparity map.

[0027] Point clouds corresponding to individual pairs are then merged into a single cloud, and outliers are removed at 145. The dense feature detection is totally data driven, without using prior face knowledge or generic faces. 150 defines dense feature computation aids, used as simplifications to the dense feature matching. This may include outlier rejection techniques (such as tensor voting), and may include area search minimization.

[0028] At 155, the final cleaned point cloud is used to form a connected surface. A face texture is acquired from a frontal image. The final result is information representative of the surface. This can be a 3-D mesh formed of triangular patches. The final result can alternately be a set of 3D points or a surface defined for example by curve splines, subdivision surfaces, or other digital surface definitions.

[0029] Further details about the operation are now provided.

[0030] Conventional stereo reconstruction has relied on the existence of multiple cameras obtaining one or more similar image pairs. Feature correspondences between those multiple image pairs are determined. The feature correspondences are subsequently triangulated to find a final three-dimensional group of points.

[0031] In an embodiment, a single camera is used to obtain multiple images, and then the images are recast as multi view stereo images. In an embodiment, the process assumes that the head is static and that the camera is moving or moved with respect to the head. While this is unlikely to be the case, this assumption provides no loss of generality; e.g., the camera can be static and the head moved, or both the camera and the head moved.

[0032] As described above, the multiple images are first analyzed at 100 to determine an initial estimate of camera pose among the images. This initial estimate uses information indicative of a face, e.g. prior face knowledge or a generic face, to carry out the estimate. It provides "sparse" information that allows the system to determine enough information to find pose and correspondence between the images.

[0033] For example, the initial estimates done with the prior face knowledge or generic face may provide information that indicates the perimeters of a face, the locations of a mask defining parts of the face, or other information. This provides information for image selection, and constrains the set of sparse features to be matched. Prior face knowledge or a generic face is used to form the sparse features, but the sparse features may be refined using data-driven optimization prior to dense features being determined.

[0034] The tracker pose estimation module investigates the images to find similar images that can be rectified against one another. The similar images comprise images which define similar poses. This hence allows selection of a subset of images to be used for reconstruction. The images are selected using both the baseline information, as well as reliably tracked feature points across multiple images.

[0035] There is always a measurement uncertainty between multiple different images. For example, as the angular baseline between a pair of images decreases, the error in the computed 3-D points is magnified. This decreased angular baseline hence increases 3-D measurement uncertainty. Less accurate 3D information can be obtained from images with smaller angular baselines between the images. As the angular

baseline increases, more accurate 3D information can be extracted—however, there is also less surface area in common between the two views, and hence fewer possible matches. Image pairs are therefore selected to balance between the measurement uncertainty, and the number of errors. For example, images with 8 to 15 degrees of angular baseline and 6 points matched across the image pair, may be preferred.

[0036] The balancing may be carried out by tracking feature points in multiple selected images. Only images which have high confidence matches (e.g., greater than 90%) between features are retained to establish feature chains. Frame pairs are maintained within the set of images if they meet the feature points and also meet a set baseline criteria. For example, the baseline criteria can be set—such as requiring at least 5 degrees of angular baseline. The feature point criterion also rejects frames that have highly inaccurate tracker pose estimates.

[0037] This sparse matching phase produces a set of images and feature points that are matched across the sequence. Matches that are supported by this feature point matching are likely to be more accurate than matches which are solely predicted by the pose tracker. Feature point matches may also cover a greater number of frames than the tracker predicted matches—and hence provide more constraints on the camera pose refinement process. These constraints may result in greater accuracy in the pose refinement at **130**.

[0038] The bundle adjustment starts with the sets of images and feature points that are matched across the image set. These have been obtained, as described above, by the feature tracking. The bundle adjustment which is carried out at **130** is an optimization technique that solves for the camera parameters and for the 3-D positions of points based on two-dimensional correspondences between sets of images. The optimized parameters may include position and orientation of the camera and 3-D structure of the 2-D feature points. The optimization may be carried out by alternating a partial solution for structure, and then a partial solution for camera pose. A computer may alternatively carry out both of these calculations until an adequate solution converges.

[0039] Bundle adjustment hence estimates the position of the camera in each image, by flip-flopping between estimating the pose of cameras and the position of points in an iterative fashion until it finally converges. The end result is a more accurate camera position as well as structure of the points. Because these are sparse “high confidence” points, it does not provide a full dense representation, but that is done in later stages.

[0040] An alternative technique may simply iteratively change the values until good values are obtained.

[0041] The 3-D locations of the matched feature points as estimated and refined bundle adjustment **130** are used in the later stages to constrain the scope of the reconstruction. These form optimized camera poses that are used in all subsequent processing stages.

[0042] Dense feature matching **140** finds more information about corresponding points among the image pairs. An unconstrained dense matching, however, can be computationally prohibitive, since it can require a full image search for each match. An unconstrained search would compare each point in each image against each point in every other image.

[0043] **150** generically represents the techniques that are used to reduce the scope of the dense feature search.

[0044] According to an embodiment, an epipolar geometry technique is used. In epipolar geometry, each correspondent item must lie along a single line that extends between the paired or clustered images. The process can be further simplified by rectifying the images, such that each epipolar line coincides with a horizontal scan line. This avoids the need to re-sample the images for each potential match.

[0045] After the rectification, corresponding points in each pair of images are found using a matching process. The prior face knowledge or generic face may be used to assist the matching process by restricting the matching to the area covered by the tracking face mask. This allows simplifying the search such that a template is extracted using a fixed window size for each pixel in one image. The template is matched along the corresponding epipolar line in the paired image.

[0046] A minimum correlation threshold and restricted disparity range suitable for faces is used to reduce the number of spurious matches. Locations with a flat correlation plot or no obvious peak are rejected. However, multiple candidate matches may be maintained to find the best match.

[0047] The result of the matching process is a disparity volume. Each (x,y,d) triplet maps a pixel (x,y) in one rectified image to a pixel (x+d,y) in a paired image.

[0048] The known poses can be triangulated to convert disparity values to three dimensional points. Each disparity pixel is transformed to its original image space using the inverse of the rectifying transform. The three-dimensional location of that match is provided by the intersection between the rays passing through the camera's optical center and the corresponding feature matches in the image plane. In reality, errors in the feature matching and camera estimates will prevent these lines from intersecting exactly. The three-dimensional point that minimizes the orthogonal distance between the rays may be used.

[0049] Another constraint may be provided by rejection of outliers in the derived structure. The three-dimensional result from the bundle adjustment process provides a more accurate, though sparse, estimate of the three-dimensional face structure. This is not sufficient to capture the subtle geometry of the face. In an embodiment, this is used to provide a constraint on the allowable three-dimensional computations in the dense reconstruction. Specifically, the computed structure should not deviate far from the bundle-adjustment-derived structure. This structure is first used to prefilter the data by converting the interpolated bundle-adjusted structure to voxels, and rejecting data at a predetermined distance from the voxels. In effect, this becomes a data optimization technique.

[0050] The voxel testing removes the gross outliers, that is those which are more than a predetermined distance from the bundle voxels. It also removes boundary artifacts that are due to inaccurate placement of the face mask. Errors in feature matching, however, may result in reconstruction noise. If the noise is uncorrelated within the views and between the views, it will appear as sparse, high frequency variations in the three-dimensional structure. Correct matches however, will be correlated between views due to the smoothness and continuity of the face structure.

[0051] Tensor voting may also be used to determine surface saliency and hence to maintain the correlation structure is tensor voting. A three-dimensional tensor voting scheme can be used to reinforce and determine the surface saliency. Tensor voting allows each 3-D point to be encoded as either a ball tensor, or a stick tensor. The information in the tensor is propagated to their neighbors via a voting operation. Neigh-

bors which have similar structure therefore reinforce each other through the tensor voting process. The amount of the structural reinforcement is influenced by the initial structural saliency. This technique recovers a surface from the cloud of points.

[0052] A good initial estimate of point normals may be preferred to blindly encoding the points as ball tensors. In an embodiment, the head is approximated by a cylinder as shown in FIG. 4A. Cylinder normals are obtained. The cylinder normals may be used as the point normal approximations.

[0053] FIG. 4B illustrates the same points unwrapped and triangulated.

[0054] In another embodiment, the system may use a 3×3 Eigensystem and may fix the normal as the first eigenvector in that Eigensystem. The remaining basis vectors may then be computed using singular value decomposition. Initial surface saliency, e.g. that defined by the difference in magnitude between the first two eigenvectors, may be set uniformly for all points.

[0055] The 3D points obtained from the bundle adjustment are very accurate but sparse estimates of the facial structure. These points are added to the tensor voting point set with boosted surface saliency. Radial basis functions may also be used to interpolate a smooth surface between the 3D points obtained from the bundle adjustment. In this embodiment, normals for the 3D bundle points are computed from the interpolated surface, to use for the tensor voting. However, the interpolated surface itself is preferably not used for the tensor voting.

[0056] After two passes of tensor voting, points with low surface saliency are removed, leaving a dense cloud of points distributed across the surface of the face.

[0057] Prior face knowledge or a generic face may be introduced in the dense reconstruction stage such that the face space is not constrained. Specifically, an embodiment may use the prior face knowledge or a generic face in the dense process to determine and reject outliers e.g., based on proximity to an existing generic face representation, but is not used to compute or modify the 3D position of reconstructed points.

[0058] The face detail is effectively captured in the three-dimensional point cloud. If the final goal is a mathematical description of the face, then the three-dimensional point cloud may be sufficient.

[0059] An embodiment uses domain knowledge to generate and texture a mesh based on the dense three-dimensional structure. The embodiment operates to unwrap the 3-D point cloud onto a two-dimensional plane via a cylindrical projection. Following the cylindrical projection, each three-dimensional point cloud has a corresponding two-dimensional map location. The two-dimensional map locations may be triangulated using Delaunay triangulation. Their connectivity information is then transferred to the three-dimensional

points, and the surface is defined according to the resulting mesh. The cylindrical unwrapping and triangulation is illustrated in FIG. 4B.

[0060] The general structure and techniques, and more specific embodiments which can be used to effect different ways of carrying out the more general goals are described herein.

[0061] Although only a few embodiments have been disclosed in detail above, other embodiments are possible and the inventors intend these to be encompassed within this specification. The specification describes specific examples to accomplish a more general goal that may be accomplished in another way. This disclosure is intended to be exemplary, and the claims are intended to cover any modification or alternative which might be predictable to a person having ordinary skill in the art. For example, many of the operations discussed herein refer to operations without using a generic face or prior face knowledge. It should be understood that these techniques alternately can be carried out using such a generic face or prior face knowledge, for some, but not all, of these techniques.

[0062] Also, the inventors intend that only those claims which use the words "means for" are intended to be interpreted under 35 USC 112, sixth paragraph. Moreover, no limitations from the specification are intended to be read into any claims, unless those limitations are expressly included in the claims. The computers described herein may be any kind of computer, either general purpose, or some specific purpose computer such as a workstation. The computer may be a Pentium class computer, running windows XP or Linux, or may be a Macintosh computer. The computer may also be a handheld computer, such as a PDA, cellphone, or laptop.

[0063] The programs may be written in C, or Java, Brew or any other programming language. The programs may be resident on a storage medium, e.g., magnetic or optical, e.g. the computer hard drive, a removable disk or media such as a memory stick or SD media, or other removable medium. The programs may also be run over a network, for example, with a server or other machine sending signals to the local machine, which allows the local machine to carry out the operations described herein.

[0064] Where a specific numerical value is mentioned herein, it should be considered the value may be increased or decreased by 20%, while still staying within the teachings of the present application, unless some different range is specifically mentioned.

What is claimed is:

1. A face reconstruction method, comprising:

analyzing a plurality of images of a face to find sparse, three-dimensional face features using prior knowledge of a face; and

using said sparse three-dimensional face features to analyze said plurality of images, to find dense three-dimensional features using a data driven approach, without using any prior knowledge.

* * * * *