ANIMATED HEADS: FROM 3D MOTION FIELDS TO ACTION DESCRIPTIONS

Jan Neumann, Cornelia Fermüller and Yiannis Aloimonos
Center for Automation Research
University of Maryland
College Park, MD 20742-3275, USA
(jneumann, fer, yiannis)@cfar.umd.edu

Abstract We demonstrate a method to compute three-dimensional (3D) motion fields on a face. Twelve synchronized and calibrated cameras are positioned around a talking person, and observe its head in motion. We represent the head as a deformable mesh, which is fitted in a global optimization step to silhouette-contour and multi-camera stereo data derived from all images. The non-rigid displacement of the mesh from frame to frame, the 3D motion field, is determined from the spatio-temporal derivatives in all the images. We integrate these cues over time, thus producing an animated representation of the talking head. Our ability to estimate 3D motion fields points to a new framework for the study of action. The 3D motion fields can serve as an intermediate representation, which can be analyzed using geometrical and statistical tools for the purpose of extracting representations of generic human actions.

1. INTRODUCTION

What does it mean to understand an action? *One understands an action if one is able to imagine performing an action with images that are sufficient for serving as a guide in actual performance.* To be able to visualize or virtualize an action in our mental theater, we have to develop a spatio-temporal action description of the object in space that is performing the action. What are the key points in figuring out the nature of action representations?

1 Action representations are view independent. We are able to recognize and visualize actions regardless of viewpoint.

N. Magnenat-Thalmann et al. (eds.), *Deformable Avatars*

2 Action representations capture dynamic information which is manifested in a long image sequence. Put simply, it is not possible to understand an action on the basis of a small sequence of frames (viewpoints).

3 Action representations are made up of a combination of shape and movement.

To gain insights on action representations we consider them in a hierarchy. First there is the image data, that is, videos of humans in action. Considering the cue of motion, then our image data amounts to a sequence of normal flow fields computed from the videos. The second kind of representations are intermediate descriptions encoding information about 3D space and 3D motion, estimated from the input (video). These representations consist of a whole range of descriptions of different sophistication encoding partially the space-time geometry, and they are view and scene dependent. Finally, we have representations encoding the characteristics of actions, and these representations are view and scene independent. The most sophisticated intermediate representation for the specific action in view that could be obtained is then a sequence of evolving 3D motion fields (also known as *range flow* (Spies et al., 2000) or *scene flow* (Vedula et al., 1999)). Acquiring this representation is no simple matter, but it can be achieved by employing a very large number of viewpoints (e.g., for a general overview about human motion modeling see (Aggarwal and Cai, 1999) and (Gavrila, 1999)).

As an example for an interesting action, we will examine facial expressions. Several image sequences of a talking and moving head were simultaneously recorded by a large number of cameras. From these image sequences a three-dimensional mesh model of the head was constructed and the trajectories of the mesh vertices in space-time, the evolving motion fields, were determined.

Due to the large number of possible applications, for example in the field of human-computer-interaction or in entertainment (e.g., "Motion Capturing"), a lot of work has been done on the creation of 3D models of faces and the synthesized and recognition of facial expressions. Most approaches made use only of a few viewpoints at a time, thus they were not utilizing all the available constraints and information. For example, (Fua and Miccio, 1999) and (Pighin et al., 1998) fitted a predefined animation model to image data from few views and (Vetter and Blanz, 1998) used a single image in an analysis-by-synthesis loop.

Other methods need complicated prior motion and face models (e.g., (Terzopoulos and Waters, 1993) and (Essa and Pentland, 1997) use a physics-based model with anatomically correct muscles) or tracking mark-

ers on the face (.e.g, (Guenter et al., 1998)) to extract the facial expressions. The difference to our approach is that we construct a full three-dimensional model without manual intervention and without relying on any prior model. The 3D motion flow on the head surface is computed directly from image derivatives, not on the basis of optical flow. Stereo and motion estimation were combined into one framework similar as in (Zhang and Kambhamettu, 2000) and (Malassiotis and Strintzis, 1997). But in their work in contrast to our approach the scene is still parameterized in the image space of the base view, whereas we use the more natural object space parameterization. By moving the representation from image to object space, the algorithm can handle arbitrary camera arrangements and can make use of robust regularization constraints on the object surface, because physical tissue deforms in a continuous and smooth manner. The use of multi-camera setups for the computation of full 3D flow has only recently become feasible due to sinking costs of image capture and computer equipment (for an example see Vedula et al., 2000).

In building scene-independent representations for facial expressions, it is essential to separate the 3D motion flow field into a component due changes of pose and a component due to the facial expression. Former approaches used simplified models such as planar models plus parallax for the head motion and affine motion models for the facial expressions (e.g., Bascle and Blake, 1998 and Black and Yacoob, 1997). By using the changing silhouettes and the rigid surface regions of the object to determine the rigid motion, we can compensate for the change in pose. After subtracting the rigid motion flow component from the full flow, we are left with the non-rigid residual motion describing the facial expression that can be analyzed or used for reanimation of other models.

2. PRELIMINARIES AND DEFINITIONS

We have established in our laboratory a multi-camera network consisting of sixty-four cameras, Kodak ES-310, providing images at a rate of up to eighty-five frames per second; the video is collected directly on disk –the cameras are connected by a high-speed network consisting of sixteen dual processor Pentium 450s with 1 GB of RAM each (Davis et al., 1999).

The camera configuration is parameterized by the camera positions $\mathbf{T}_k$, the rotation matrices R_k that relate the camera coordinate system to the fiducial system, and the intrinsic camera parameters K_k (bold-face letters denote vectors, small letters scalars, and large letters matrices). The calibration is done using images of a large calibration object. In

Figure 1 Calibrated Camera Setup with Example Input Views

the following we assume that the images have already been corrected for radial distortion. The image formation process is described by the conventional pinhole camera model, where the point $\mathbf{P}$ in fiducial world coordinates is related to its projection $\mathbf{p}_k$ in camera k as follows ($\hat{\mathbf{z}} = [0\,0\,1]^\top$):

$$\mathbf{p}_k = K_k \frac{R_k(\mathbf{P} - \mathbf{T}_k)}{\hat{\mathbf{z}} \cdot R_k(\mathbf{P} - \mathbf{T}_k)} \tag{1}$$

The head surface is approximated by a closed mesh with vertices $\mathbf{V}_i$ and triangular facets $\mathbf{F}_j$. The world coordinates of $\mathbf{V}_i(t) = [x_i(t), y_i(t), z_i(t)]$ are dependent on time t. Since we formulate the structure and motion estimation in object space, the image information needs to be sampled in regular patterns on the mesh surface instead of in regular patterns on the images. Therefore, a set of regularly spaced sampling points is associated with each triangle. The number of sampling points is dependent on the visible area of the triangle in the different cameras.

It is assumed that the head is the only moving object in all the image sequences, although this assumption is not essential and can be eliminated by applying the algorithm in turn to each independently-moving object. The following sections describe the algorithm that computes the spatio-temporal representation of the moving and talking head (from now on called the "object"):

- Section 3: Motion-based segmentation of the input images to locate the moving object, compute its silhouettes, and initialize the deformable 3D mesh.

- Section 4: Multi-camera stereo refinement of the deformable mesh where the search space is constrained by the silhouettes.
- Section 5: Computation of the 3D motion field on the mesh surface from image derivatives based on the normal flow constraint.

3. IMAGE SEGMENTATION

We incrementally construct an image of the background by modeling the temporal evolution of the changing foreground pixels and the static background pixels. The magnitude of the temporal image derivatives and image statistics such as mean and variance are computed for each pixel on ten consecutive frames in the sequence and then used to segment the image into fore- and background. We integrate information over time to make the segmentation more robust by applying order-statistic filters over small spatio-temporal volumes. After the initial segmentation , we intersect the cone-shaped spaces formed by reprojecting the convex hulls of the head silhouettes into space. The intersection is a convex approximation of the head and it defines the initial 3D surface mesh. The mesh is now back-projected into each image and the segmentation is refined by fitting the mesh to all silhouette contours simultaneously.

4. MULTI-CAMERA STEREO ESTIMATION

Using only information from silhouettes, it is not possible to compute more than the visual hull (Laurentini, 1994) of the object in view. Therefore, to refine our 3D surface estimate of the object, we adapt the vertices of the mesh to optimize the correlation between corresponding image regions in the different camera views. The search range for the vertex positions is constrained by the displacement boundaries computed in the silhouette estimation step in Section 3. To determine the visibility of each triangle, a z-buffer algorithm computes the index of the closest triangle patch for each pixel location. Next, a regular sampling point pattern is assigned to each mesh triangle as described before in Section 2, so that the sampling density of the closest image is about one projected sampling point per pixel.

We optimize orientation and position of each triangle by displacing each triangle vertex along the surface normal direction of the mesh and maximizing a similarity criterion among the triangle projections. The criterion to be optimized is the normalized cross-correlation between the projections of each triangle into all the cameras in which the triangle is visible (we denote this set of cameras as the set of "visible cameras"). For all combinations of normal displacements of the three vertices we compute the 3D coordinates of the sampling points on the triangle sur-

face and project the sampling points into all the visible cameras. The image brightness of a projected sampling point is determined by bilinear interpolation. The cross-correlation is now computed between the corresponding image brightness samples for all pairs of cameras that mutually see the triangle. We combine the correlation scores from all the camera pairs by taking a weighted average with the weights depending on the angle between camera plane and triangle plane.

The pairwise scores between all the cameras are also used to correct the visibility information. If a bimodal distribution of high and low correlation scores can be detected, then it is possible to estimate which cameras are visible and which are not, and the occluded cameras can be excluded from the score. For each vertex we collect the normal displacements corresponding to the highest correlation score for each of the surrounding triangles and determine the final normal displacement subject to global smoothness and rigidity constraints which have been added to regularize the solution.

5. MOTION ESTIMATION

Following the description of the photometric properties of a surface in space in (Horn, 1986) and (Vedula et al., 1999), the head surface is assumed to have Lambertian reflectance properties, thus the brightness intensity of a pixel $\mathbf{p}_k$ in camera k is given by

$$I(\mathbf{p}_k; t) = -c_k \cdot \rho(\mathbf{P}) \cdot [\mathbf{n}(\mathbf{P}; t) \cdot \mathbf{s}(\mathbf{P}; t)] \tag{2}$$

with an albedo $\rho(\mathbf{P})$ that is constant over time ($d\rho/dt = 0$) and where c_k is the constant that describes the brightness gain for each camera, $\mathbf{n}$ is the normal to the surface at $\mathbf{P}$, and $\mathbf{s}$ the direction of incoming light. Taking the derivative with respect to time on both sides, we get the following expression for the change of the image brightness $I(\mathbf{p}_k)$ at pixel location $\mathbf{p}_k$ in camera k:

$$\frac{dI(\mathbf{p}_k)}{dt} = \nabla I(\mathbf{p}_k) \cdot \frac{d\mathbf{p}_k}{dt} + \frac{\partial I(\mathbf{p}_k)}{\partial t} = -c_k \cdot \rho(\mathbf{P}) \cdot \frac{d}{dt}[\mathbf{n}(\mathbf{P}; t) \cdot \mathbf{s}(\mathbf{P}; t)] \tag{3}$$

Since our sequences were recorded with a frame rate of 60 Hz and under fixed illumination, we can assume that $\frac{d}{dt}[\mathbf{n} \cdot \mathbf{s}] = 0$, and we end up with the well-known *normal flow constraint* equation.

$$-\frac{\partial I(\mathbf{p}_k)}{\partial t} = \nabla I(\mathbf{p}_k) \cdot \frac{d\mathbf{p}_k}{dt} \tag{4}$$

This equation gives us one constraint per measurement, we can only determine the component of the optic flow that is normal to the image gradient, the normal flow. The estimation of the tangential flow

along the iso-brightness contour is ill-posed. Regularizing the problem by imposing image-based smoothness conditions on the solution to equation (4) introduces artifacts at depth discontinuities and biases due to inhomogeneous gradient distributions (Fermüller et al., 2000).

Each normal flow vector in an image constrains the projection of the 3D motion flow to lie along a line parallel to the iso-brightness contour in the image, the normal flow constraint line. Thus the 3D motion flow vector has to lie on the plane defined by the normal flow constraint line and the optical center of the camera. The component of the 3D motion along the iso-brightness contour on the object surface is not recoverable. This is the aperture problem revisited in 3D. Nevertheless, if we assume that neighboring patches on the surface will move in an elastic manner, we can impose smoothness constraints on the motion of neighboring points. This smoothness assumption is physically justified as long as our mesh model has the same topology as the object in view, because nearly all real materials deform elastically when strain is applied.

The mesh representation of the head defines a correspondence map between the cameras, and the full 3D motion flow at each mesh vertex is determined by combining the information from all the sampling points of the triangles neighboring the mesh vertex. To relate image derivatives and 3D motion flow using the normal flow constraint, we have to determine the Jacobian of the image formation equation (1) (R_3 is third row of matrix R and $K, R, \mathbf{T}$ refer to the calibration parameters of camera k):

$$\frac{d\mathbf{p}_k}{dt} = \frac{\partial \mathbf{p}_k}{\partial \mathbf{P}_k}\frac{\partial \mathbf{P}_k}{\partial t} = \frac{\partial P}{\partial t} K \frac{R(\mathbf{P}-\mathbf{T})}{R_3(\mathbf{P}-\mathbf{T})} = \left(\frac{KR - \mathbf{p}_k R_3}{R_3(\mathbf{P}-\mathbf{T})}\right)\frac{\partial \mathbf{P}}{\partial t} \tag{5}$$

The derivative images are sampled at all locations where the sampling points associated with each triangle are visible. Let a given triangle of the mesh be defined by the vertices $\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3$, then for each sampling point $P = \sum_{j=1,2,3} \lambda_j \mathbf{V}_j$ of this triangle we get the following constraint equation for each measurement:

$$-\frac{\partial I(\mathbf{p}_k)}{\partial t} = \sum_{j=1,2,3} \lambda_j \left(\nabla I(\mathbf{p}_k) \cdot \frac{KR - \mathbf{p}_k R_3}{R_3(\mathbf{P}-\mathbf{T})}\right)\frac{\partial \mathbf{P}_j}{\partial t} \tag{6}$$

There is one equation per sampling point per visible image. To integrate these constraints, we stack these equations to form the $m \times n$ matrix L where m is the number of sampling points over all the triangles and their projections into all the visible cameras and n the number of vertices of the mesh times the three spatial dimensions. The matrices for the models presented are on the order of $100\,000 \times 3000$. To regularize the solution we add smoothness constraints as extra rows to L.

Since it is computationally infeasible to solve this large system directly, we form the normal equations of the over-constrained system and solve them with a preconditioned conjugate gradient method with either the motion field of the previous frame or the solution to a rigid motion approximation as starting vectors. The second choice worked very well to initialize the optimization, because most parts of a human head move rigidly. The magnitude of the residual non-rigid flow is used to segment the mesh into rigidly and non-rigidly moving areas. This enables us to separate the motion field into two parts, one due to the change of pose and one due to the expression on the face.

6. RESULTS

For our experiments we used eleven cameras placed in a dome-like arrangement around the head of a person that was expressing surprise (Figure 1). After the initial structure estimation stage of our algorithm, we are able to synthesize texture-mapped views of the head from arbitrary viewing directions (Figures 2a-2c). The textures, coming always from the least oblique camera with respect to a given triangle, were not blended together to demonstrate the good agreement between adjacent texture region boundaries. This demonstrates that the spatial structure of the head was recovered very well.

The 3D motion flow field for the current frame is computed and used to propagate the mesh to the next frame. The propagated mesh is refined by new stereo and silhouette data, before the next 3D motion flow field is computed, and the process is repeated. The 3D motion field shown in (Figures 2d-2fl) was computed by integrating the 3D flows of frames 40 to 45.

The rigid motion flow was computed by parameterizing the 3D motion flow vectors by the instantaneous rigid motion $\partial \mathbf{P}/\partial t = \mathbf{v} + \boldsymbol{\omega} \times \mathbf{P}$, where $\mathbf{v}$ and $\boldsymbol{\omega}$ are the instantaneous translation, and rotation (Horn, 1986). This parameterized flow field was then fitted to the image derivative information in the images. By subtracting the rigid motion flow from the full flow, we extract the non-rigid flow. It can be seen that the rigid motion part (the turning of the head to the upper left) is recovered well, as the magnitude of the residual non-rigid flow on the rigid part of the head (e.g., forehead, nose and ears) in Figure (2e) is significantly smaller than the full flow in Figure (2d).

The non-rigid motion is also computed accurately, as we can easily see in the close up of the mouth region (Figure 2f), how the mouth opens, and the skin of the jaw stretches recedes. Animations of the re-

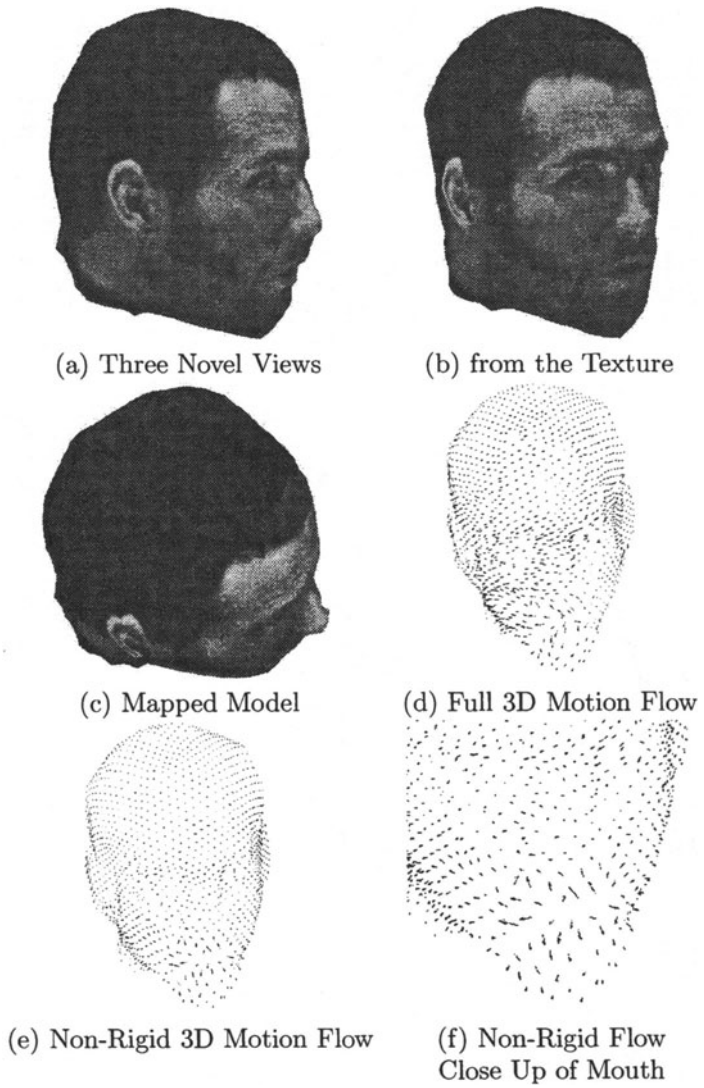

Figure 2 Results of 3D Structure and Motion Flow Estimation

covered model and flow fields can be found at the following web address: *http://www.videogeometry.com/TalkingHeads.*

7. CONCLUSION AND FUTURE WORK

We presented an algorithm that computes an accurate spatio-temporal description of a non-rigidly moving human head. The description consists of the spatio-temporal trajectories of the mesh vertices, the evolving motion fields.

To see how these motion fields can be used, let us now consider the mapping from the 3D motion fields to the scene independent action representations. This mapping should be such that it extracts from a specific action quantities of a generic character common to all actions of the same type. These quantities most probably take the form of spatio-temporal patterns in four dimensions.

One way of obtaining such patterns is to perform statistics on a large enough sample (e.g., Reynard et al., 1996). Considering, a particular action (e.g., talking or dancing), we can obtain data in the multi-camera laboratory described before for a large number of individuals. In each case we can obtain a 3D motion field and thus are able to build up a large data base of 3D motion fields. To this database a number of statistical techniques, such as principal component analysis, can be applied to reduce the dimensionality of the space and describe it with a small number of parameters. Another way of obtaining these patterns would be to study invariances related to symmetry, and geometric quantities in space-time (e.g., angles, velocities, accelerations, periodicity, etc. (Bottema and Roth, 1979)).

In our future work, we will apply the above mentioned statistical and geometrical methods to the evolving 3D motion fields and try to extract the action representations. To improve the presented algorithm we plan to incorporate explicit visibility updating into the stereo part of the algorithm and include further information such as range flow constraints (see Spies et al., 2000) between the consecutive stereo reconstructions.

References

Aggarwal, J. and Cai, Q. (1999). Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440.

Bascle, B. and Blake, A. (1998). Separability of pose and expression in facial tracing and animation. In *ICCV98*, pages 323–328.

Black, M. and Yacoob, Y. (1997). Recognizing facial expressions in image sequences using local parameterized models of image motion. *IJCV*, 25(1):23–48.

Bottema, O. and Roth, B. (1979). *Theoretical Kinematics*. North-Holland.

Davis, L., Borovikov, E., Cutler, R., Harwood, D., and Horprasert, T. (1999). Multi-perspective analysis of human action. In *Proc. of Third*

International Workshop on Cooperative Distributed Vision, Kyoto, Japan.

Essa, I. and Pentland, A. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. PAMI*, 19(7):757–763.

Fermüller, C., Pless, R., and Aloimonos, Y. (2000). The Ouchi illusion as an artifact of biased flow estimation. *Vision Research*, 40:77–96.

Fua, P. and Miccio, C. (1999). Animated heads from ordinary images: A least-squares approach. *Computer Vision and Image Understanding*, 75(3):247–259.

Gavrila, D. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98.

Guenter, B., Grimm, C., Wood, D., Malvar, H., and Pighin, F. (1998). Making faces. In *Proc. of SIGGRAPH*, pages 55–66.

Horn, B. K. P. (1986). *Robot Vision*. McGraw Hill, New York.

Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *IEEE Trans. PAMI*, 16(2):150–162.

Malassiotis, S. and Strintzis, M. (1997). Model-based joint motion and structure estimation from stereo images. *Computer Vision and Image Understanding*, 65(1):79–94.

Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., and Salesin, D. (1998). Synthesizing realistic facial expressions from photographs.

Reynard, D., Wildenberg, A., Blake, A., and Marchant, J. (1996). Learning dynamics of complex motions from image sequences. In *ECCV96*, pages I:357–368.

Spies, H., Jaehne, B., and Barron, J. (2000). Dense range flow from depth and intensity data. In *ICPR00*.

Terzopoulos, D. and Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. PAMI*, 15(6):569–579.

Vedula, S., Baker, S., Rander, P., Collins, R., and Kanade, T. (1999). Three-dimensional scene flow. In *ICCV99*, pages 722–729.

Vedula, S., Baker, S., Seitz, S., and Kanade, T. (2000). Shape and motion carving in 6d. In *CVPR00*, pages II:592–598.

Vetter, T. and Blanz, V. (1998). Estimating coloured 3-d face models from single images: An example-based approach. In *ECCV98*, pages 499–513.

Zhang, Y. and Kambhamettu, C. (2000). Integrated 3d scene flow and structure recovery from multiview image sequences. In *CVPR00*, pages II:674–681.