Correspondence and Affine Shape from two Orthographic Views: Motion and Recognition

Amnon Shashua

Abstract

The paper presents a simple model for recovering affine shape and correspondence from two orthographic views of a three–dimensional object. The paper has two parts. In the first part it is shown that four corresponding points along two orthographic views, taken under similar illumination conditions, determine affine shape and correspondence for all other points. In the second part it is shown that the scheme is useful for purposes of visual recognition by generating novel views of an object given two model views in full correspondence and four corresponding points between the model views and the novel view. It is also shown that the scheme can handle objects with smooth boundaries, to a good approximation, without introducing any modifications or additional model views.

1 Introduction

Structure from motion (SFM) and visual recognition are intimately related. Recovering the structure of a moving three-dimensional (3D) object from its changing 2D image is dual to the problem of identifying images of an object viewed from a variety of vantage points, as instances of the same 3D object. Both require an understanding of the relationship between the 3D world and its 2D projections, both start with the same input and both work with essentially the same ingredients: 3D structure of the object, motion or viewing transformation applied to the object, and the pointwise correspondence between two or more views of the object.

In SFM one generally wants to recover information that was lost in the course of projection from 3D to 2D. This includes the 3D Euclidean structure of the object and the 3D motion transformation from one time instance to the next. Visual recognition confronts the same issues but in a more implicit manner. Rather than recovering 3D information, one is more concerned in factoring it's effects out, i.e. the effect of shape and viewing transformation, thereby reducing all views of an object to a canonical view (or set of views) that represents the object.

Previous approaches to 3D interpretation traditionally assume that correspondence between 2D views is known, or can be measured independently [4, 55, 23, 3, 21, 31]. Under perspective projection it has been shown that two views undergoing infinitesimal motion are, in principal, sufficient to recover shape and motion [41, 30, 55, 54, 36], however the process in inherently susceptible to noise [17, 2, 46, 11]. Under orthographic projection, it has been shown that at least three views, undergoing general motion, are required to recover the same information [48, 49, 25, 8, 47]. In object recognition, the approach that seems most relevant to known results from structure from motion is the *alignment* approach [16, 19, 20, 50, 26]. Under this framework it has been shown [50, 26] that a 3D model together with a small number of corresponding points are sufficient for predicting novel views of the rigid object, and recently that shape information can be represented [51], or approximated [18], by having instead a set of 2D views of the object.

The approach to SFM in this study is different from most past approaches in that it is guided by a specific goal — performing visual recognition. This implies that information to be recovered from the changing 2D image should be no more than what is necessary to perform visual recognition. Instead of recovering Euclidean shape[1] and 3D motion parameters, the emphasis here is to recover affine shape[2] and full correspondence between two orthographic views, given limited informa-

---

[1] 3D coordinates relative to a Cartesian frame aligned with the viewer's coordinate system and with the line of sight.

[2] 3D coordinates relative to a frame defined by an arbitrary set of four non-coplanar points on the object.

tion regarding motion parameters — information that is captured by having four corresponding points between the two frames.

The reason for the emphasis on recovering affine shape is twofold. It will be shown that affine shape recovered from the correspondence between two model views is sufficient for purposes of recognition — one can generate novel views (excluding occlusion) of the object undergoing arbitrary 3D affine transformations, given four corresponding points with the novel view. Furthermore, affine shape seems to play an important role in the perception of kinetic depth displays, even in cases where Euclidean shape can theoretically be recovered, as suggested in [45].

The emphasis on solving the correspondence problem is inspired from recent developments in visual recognition using alignment [51, 43] and Radial Basis Functions [18] indicating that establishing correspondence between two or more views is a major step towards ameliorating the effects of changing view position and illumination conditions. The main new results presented in this study include the following:

- Four corresponding points along two orthographic views, taken under similar illumination conditions, together with the instantaneous brightness measurements are sufficient to completely determine, without regularizing assumptions, correspondence and affine shape along all other points in the image.

- The information carried by the four corresponding points can be succinctly represented by a 2D affine transformation that serves as a constraint line in correspondence space. The scale factor associated with the affine displacement vector is a shape parameter representing the relative deviation, along the line of sight, of an object point from a reference plane defined by three of the corresponding points. This result is new in its algebraic aspect; the concept of representing affine shape as a deviation from a reference plane was recently introduced by Koenderink and Van-Doorn [28].

- The computational study suggests that the measurement of motion starts by setting up a frame of reference determined by a small number of salient, unambiguously matched, features. The frame provides a *nominal* motion, which is exact for planar surfaces, and which 'pulls' or 'captures' all other points in that frame. The remaining *residual* motion is later refined by use of local spatio-temporal detectors that are tuned along a known direction which is determined by the frame of reference.

- The result that correspondence can be recovered from two views under similar illumination conditions suggests that small changes of view position, can be factored out in the course of recognition, using only a single picture of the object as a model. Another result is that affine shape recovered from

2

the correspondence between two model views can be used to generate novel views of the object undergoing arbitrary 3D affine transformations, given four corresponding points with the novel view. It is also shown that this result applies to objects with smooth boundaries, to a good approximation, without introducing additional model views. (Objects with smooth boundaries, such as ellipsoids or spheres, are more complex because the object's boundary contour is not projected from fixed contours on the object [7, 27]).

The remainder of this section presents the results concerning establishing correspondence and affine shape from two orthographic views (the first three items above). Section 2 puts these results in the context of visual recognition (fourth item above).

1.1   Shape and Correspondence from 2 Views

We assume orthographic views at time instances, t_1 and t_2, are taken of a surface in 3D space. We assume the convention that the 3D Cartesian frame is aligned with the $x - y$ axis in image space, and that the z axis is along the viewer's optical axis. Furthermore, without loss of generality, we assume that the origin of the 3D frame is aligned with the point $(0, 0)$ in the image plane. The following notation is used. Let P be a point in 3D space at time t_1, and $p = \sigma[P]$ be its orthographic projection onto the image plane. Let P' be the location of the point P at time t_2, and $p' = \sigma[P']$ be the image space coordinates of P'. We therefore refer to the pair p and p' as *corresponding* points. Let $op = p - o$ denote the vector from the point o to p, i.e. op represents the coordinates of p with respect to a new origin located at point o. Similarly $OP, O'P', o'p'$ denote the vectors from O to P, from O' to P' and from o' to p', respectively. A point p will be referred to as *privileged* if its corresponding point p' is given as input.

Let O, P_1, P_2, P_3 be four non-coplanar *reference* points[3] on an object of interest in 3D. Taking O to be the origin, we obtain a 3D affine coordinate frame, and therefore, any point P on the object can be represented in the affine coordinate frame with its associated set of coordinates b_1, b_2, b_3 in the following way:

$$OP = \sum_{j=1}^{3} b_j (OP_j)$$

The crucial point is that the b's are invariant with respect to linear transformations applied to the equation above (which correspond to affine transformations in space that include rotation, translation, scaling and shearing of the object).

---

[3] The term *reference* point is adopted from projective geometry (see [42]).

Let the object undergo an arbitrary affine transformation in space, and let O', P_1', P_2', P_3' and P' be the new space locations of the affine coordinate frame and the point of interest P. We therefore have:

$$O'P' = \sum_{j=1}^{3} b_j (O'P_j').$$

Under orthographic projection, we have the following relation between the image coordinates of the affine frame in both views, and the image coordinates of the point of interest in both views:

$$op = \sum_{j=1}^{3} b_j (op_j) \tag{1}$$

$$o'p' = \sum_{j=1}^{3} b_j (o'p_j'). \tag{2}$$

The four equations in formulas 1,2 combine together shape, i.e. affine coordinates, projected motion, i.e. motion of four points, and correspondence. Therefore, given the projected motion, captured by four corresponding points, and the affine coordinates we can immediately obtain correspondence as well. Also, given the correspondence $p \longleftrightarrow p'$ we have 4 equations for 3 affine coordinates which also shows that a 'view and a half' is sufficient for recovering affine shape (see also [35, 51]). Note also that formula 1 provides two equations for solving for the affine coordinates — the third equation has been lost because of the projection from 3D to 2D.

We can compensate for the loss of the third equation by producing an equation directly from the changing brightness[4]. We assume that both views are taken under identical illumination conditions, namely, that brightness change is induced purely by motion and not by photometric effects of changing viewing angle or angle between light sources and surface orientation. In other words, we assume that the brightness of an image point p is equal to the brightness of its corresponding point p' in the second view (Horn and Schunk [23]). By further assuming that the motion is infinitesimal (an assumption that will be relaxed later on), we obtain from the expansion of the total derivative of brightness at p a linear approximation to the change of brightness due to motion, known as the *constant brightness equation* [23]:

$$\nabla I \cdot \mathrm{v} + I_t = 0$$

---

[4]The term 'brightness' has different meanings in vision literature. Here it is referred to the raw image intensities (term adopted from Horn [22]).

where $\mathrm{v} = p' - p$ is the unknown displacement vector, ∇I is the gradient at point p in the image of the first view, and I_t is the temporal derivative at p. The constant brightness equation provides only one component of the displacement vector v, the component along the gradient direction, or normal to the isobrightness contour at p. This 'normal flow' information, provided by the changing brightness, is sufficient to uniquely determine the affine coordinates b_j at p, as shown next. By subtracting equation 1 from equation 2 we get the following relation:

$$\mathrm{v} = \sum_{j=1}^{3} b_j \mathrm{v}_j + (1 - \sum_j b_j) \mathrm{v}_o \qquad (3)$$

where $\mathrm{v}_j \; j = 0, .., 3$ are the known displacement vectors of the privileged points. By substituting equation 3 in the constant brightness equation we get a new equation in which the affine coordinates are the only unknowns:

$$\sum_j b_j [\nabla I (v_j - v_o)] + I_t + \nabla I v_o = 0. \qquad (4)$$

Equations 1, and 4, provide a complete set of linear equations (ignoring singular cases) to solve for the affine coordinates from which, in return, we obtain correspondence. We have therefore proven the following '4pt + brightness' proposition:

Proposition 1 (4pt + brightness) *Two orthographic images of a shaded 3D surface with four clearly marked reference points, admit a complete set of linear equations representing the affine coordinates of all surface points (excluding singular cases), provided that the surface is undergoing an infinitesimal affine transformation and the two orthographic images are taken under identical illumination conditions.*

Comments

Rigidity: note that rigidity is not required for solving for affine coordinates and correspondence. If correspondence is the main concern, say for model building [51, 43, 18, 6], then by assuming the transformation between the two views to be any linear transformation, allows one to tolerate certain non-rigid transformations, as long as the field of view is sufficiently small. This may also be relevant for a surface undergoing a rigid transformation but viewed under situations that do not fully meet the requirements of the orthographic projection model. This notion, however, is not pursued further here.

Identical Illumination Conditions: the assumption of identical illumination conditions is a useful approximation for a Lambertian surface under multiple light

sources or hemispherical illumination. In those cases the change of brightness due to motion in space is much larger than the change in brightness induced by photometric effect, such as changing viewing direction or illumination. The assumption holds exactly for an object rotating around the vertical axis under hemispherical illumination, a situation which is quite common in natural environments. (See also [24, 53, 34] for quantitative and experimental analysis). Local photometric effects can also be ameliorated to some degree by applying a linear operator, such as the Laplace operator, to the brightness values, prior to using the constant brightness equation (Bergen and Adelson [9]).

1.2   Constraint Lines in Correspondence Space

The system of equations leading to Proposition 1 can be decomposed into two constraint lines intersecting at p' for any given point p. One constraint line comes directly from the constant brightness equation: a line passing through the point $p - \frac{I_t}{|\nabla I|}\nabla I$ in direction perpendicular to the direction of the gradient ∇I at point p. The second constraint line can be derived from equations 1 and 2 as shown below.

We rewrite equations 1 and 2 in matrix form: Let M be a 2×3 matrix whose column vectors are op_1, op_2, op_3, and similarly M' has $o'p'_1, o'p'_2, o'p'_2$ as column vectors. We therefore have: $op = Mb$ and $o'p' = M'b$. Since the system $op = Mb$ is underdetermined, then the solution b is determined only up to an element of the null space of M, namely, for every solution $\tilde{r}$ the vector $\tilde{r} + \alpha\tilde{s}$ is also a solution, where α is a scale factor and $M\tilde{s} = 0$. We can substitute b in the system $o'p' = M'b$ by $\tilde{r} + \alpha\tilde{s}$ and obtain the following constraint line equation:

$$p' = o' + M'\tilde{r} + \alpha M'\tilde{s} = r + \alpha s. \qquad (5)$$

Note that r depends on p whereas s is fixed for all points, therefore the constraint lines passing through all points in the image of the moving surface are parallel to each other. The unknown parameter α can be found by using the first constraint line whenever the gradient is non–vanishing and is not perpendicular to s, i.e. s is not in the direction of the isobrightness contour at p (see Fig. 1). We have therefore proven the following proposition:

Proposition 2 *A constraint line in correspondence space can be recovered from four corresponding points along two orthographic views of an object undergoing an arbitrary affine transformation in space.*

Different versions of this result have been proposed in the past. Huang and Lee [25] and Basri [6] derive the same constraint line (which is different from the one presented here) by different approaches. Huang and Lee assume a rigid transformation
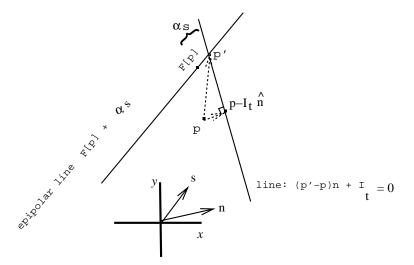
6

Figure 1: Two constraint lines that intersect at the corresponding point p'. The vector n is the normal component of the displacement vector, $n = \nabla I$ and $\hat{n} = \frac{\nabla I}{|\nabla I|}$. The vectors n, r and the scalars α, I_t are a function of the location p. The vector s is fixed for all points and can be determined only up to a scale factor.

and use that as an algebraic constraint to derive a constraint line. Basri's derivation is based on the result, originally developed in [51], that all views of an object undergoing an affine transformation in space are spanned by a linear combination of two views. This also shows that rigidity is not required for obtaining the constraint line. Koenderink and Van-Doorn [28] and Lamdan and Wolfson [29] derive a particular case of equation 5, the case where $\tilde{r}_3 = 0$.

The displacement vector $p' - p$ varies with the 3D coordinates of P and with the affine transformation applied to the object in space. Huang and Lee [25] have shown that the contribution of depth and motion cannot be decoupled from two orthographic views. The following result shows that a particular case of equation 5 can be realized by a 2D affine transformation defined by the four privileged points, to which α is a fixed function of the 3D affine coordinates of P, namely, is motion invariant.

Proposition 3 *Four corresponding points, orthographically projected from four reference points in space, determine a 2D affine transformation A,w that represent a constraint line in correspondence space, $o'p' = A(op) + w + \alpha w$, where α is a fixed function of the affine coordinates of P and is independent of the object's motion.*

Proof: The four corresponding points define three, non-collinear, corresponding vectors $op_j \longleftrightarrow o'p'_j$ $j = 1, 2, 3$. Because of non-collinearity of the vectors, there

7

exists a unique 2D affine transformation, A,w that aligns the corresponding vectors:

$$o'p'_j = A(op_j) + w, \qquad j = 1, 2, 3 \tag{6}$$

where A is a 2×2 matrix and w is a 2×1 vector. Applying the affine transformation to an arbitrary point p, yields the following result:

$$A(op) + w = A(\sum_j b_j(op_j)) + w = \sum_j b_j(o'p'_j - w) + w = o'p' + (1 - \sum_j b_j)w. \tag{7}$$

Equation 1 was used in the second term, equation 6 in the third term and equation 2 in the last term. After rearrangement we get:

$$p' = [A(op) + o' + w] + (\sum b_j - 1)w. \tag{8}$$

▯

The proposition contains two statements: the first is that under an affine coordinate frame one can derive a constraint line from four corresponding points such that the remaining degree of freedom α depends only on shape, i.e. is motion invariant. The second statement is that all of the above is captured by a 2D affine transformation derived directly from the four corresponding points.

The first statement is not new and has been introduced recently by Koenderink and Van-Doorn [28] by geometrically constructing a constraint line for which $\alpha = b_3$. An algebraic version of their result, which also shows that it is a particular case of equation 5 with $\tilde{r}_3 = 0$ is given in appendix 1. Koenderink and Van-Doorn have also derived the geometrical equivalent of α showing that it represents the relative deviation, along the line of sight, of P from the plane passing through the three reference points — thereby showing that shape is recovered up to depth scaling and shear.

The geometrical equivalent of α follows directly from the 2D affine representation of the constraint line by noticing that $\sum_j b_j = 1$ for every point P that is coplanar with the three reference points. Therefore, the transformation $A(op) + w$ accounts for the projected motion of a plane — a result well known in projective geometry [42] — and that $\alpha = \sum_j b_j - 1$ represents the deviation of P from that plane. The geometrical interpretation, much of which was described earlier in [28], can be summarized in the next proposition.

The following notations are added. The plane passing through P_1, P_2, P_3 is referred to as the *reference plane*. The point $\tilde{P}$ is the orthographic projection (along the line of sight) of the point P onto the reference plane. The point $\tilde{P}'$ is the new location of $\tilde{P}$ following an affine transformation T in space.

Proposition 4 *The 2D affine transformation defined in Proposition 3 admits the following interpretation: the affine vector w is the projection of the vector $\tilde{O}' - O'$*

8

onto the image plane and is perpendicular to the xy projection of the rotation axis of the transformation T. If T is a similarity transformation (rotation, translation and scale), then α associated with the point p is equal to:

$$\alpha = \frac{z - \tilde{z}}{\tilde{z}_o - z_o}$$

where $z, \tilde{z}, z_o$ and $\tilde{z}_o$ are the depth values of $P, \tilde{P}, O$ and $\tilde{O}$, respectively.

Proof: See appendix 2. $\square$

The affine shape parameter α provides, therefore, shape modulo translation in depth, depth scaling and shear [28]. Translation in depth is unavoidable in orthographic projection, depth scaling comes from the distance, $\tilde{z}_o - z_o$, between the reference point O and the reference plane, and shear comes from the distance, $z - \tilde{z}$, between object points and the reference plane, whose orientation is unknown. Therefore, different sets of four reference points are associated with different orientations of the reference plane and, therefore, give rise to different affine shape.

The question that is dealt with next is whether shape modulo depth scale and shear is the most one can obtain from two orthographic views. It has been shown by Ullman [49] that when the two views are separated by an infinitesimal angle rotation, then shape can be recovered up to an overall depth scaling. The depth scaling proposition holds also for planar objects but with an added ambiguity, namely, the orthographic velocity field determines exactly two solutions, each up to a depth scaling [49]. The depth scaling proposition no longer holds under finite angle transformations, as shown next, and the best one can achieve is shape up to depth scale and shear, namely affine shape. In order to eliminate the shear component from the affine shape one has to uniquely recover the equation of the reference plane, up to an overall depth scaling, and therefore the more general question is whether the depth scaling proposition holds for planar objects under finite angle transformations. The result shown below is that the parameters of the appropriate constraint line and the equation of the depth scaled plane admit a linear one parameter family of solutions. Therefore, one cannot possibly recover the plane up to a depth scaling from only two orthographic views separated by a finite angle transformation.

Proposition 5 *The constraint line parameters B, s and shape parameters a, b describing the motion of a planar object $z = ax + by + 1$, can be determined, up to a linear one parameter family of solutions, from four corresponding points along two orthographic views of the plane undergoing an affine transformation in space.*

Comments. The equation $ax + by + 1$ determines the depth $\tilde{z}$ of points on the plane up to a translation and depth scaling, i.e. $\frac{\tilde{z} - z_o}{\tilde{z}_o - z_o} = ax + by + 1$ where z_o is the

9

depth of the moving origin, $\tilde{z}_o$ is the depth of the point where the plane intersects the line of sight, and x, y are coordinates relative to x_o, y_o. Therefore, if a, b can be determined uniquely, then by subtracting $ax + by + 1$ from the affine shape α we obtain shape of a non-planar object up to depth scaling. The proposition states that one cannot determine a, b uniquely from just two views.

Proof: We subtract the corresponding point $o \longleftrightarrow o'$ from both views and the remaining three corresponding points are used to determine the parameters of the following constraint line:

$$p'_j = Bp_j + (ax_j + by_j + 1)s \qquad j = 1, 2, 3$$

where B is a 2×2 matrix and s is a 2×1 vector. Given the affine transformation defined by $p'_j = Ap_j + w$ we have from Proposition 3 that $s = \rho w$ for some constant ρ and that $B - A$ is a projection matrix $\mu[ww^t]$, for some constant μ. We have therefore,

$$p'_j = (A + \mu[ww^t])p_j + (\rho ax_j + \rho by_j + \rho)w \qquad j = 1, 2, 3,$$

which is reduced to,

$$0 = \mu(w^t p_j)w + (\rho ax_j + \rho by_j + \rho - 1)w$$

from which we get the following linear system of three equations for the four unknowns $\mu, \rho a, \rho b, \rho$:

$$1 = \mu(w^t p_j) + \rho ax_j + \rho by_j + \rho.$$

These equations are linearly independent as long as the three points are not collinear. The system is underdetermined with any number of corresponding points, because any additional point must be coplanar with the three reference points and therefore is a convex combination of these points.□

Propositions 3,4 and 5 put together show that the 2D affine transformation, recovered directly from four corresponding points, represent all the information possible from two orthographic views. $A(op) + o' + w$ accounts for the projected motion of all points P that are coplanar with P_1, P_2, P_3, and the residual for non-coplanar points is simply a vector along w whose length relative to w represents the shape of the object up to depth scaling and shear. The next proposition shows that the magnitude of the residual motion for non-coplanar points is bounded from above by the depth variation between the surface and the reference plane.

Proposition 6 *Let V_1, V_2 be two orthographic views produced by a rigid transformation, and let $\tilde{V}'_1$ be the view V_1 followed by the 2D affine transformation of Proposition 3. The remaining distance between points $\tilde{p}'$ in $\tilde{V}'_1$ and their corresponding points p' in V_2 is bounded by $\mid z - \tilde{z} \mid$ the relative depth between P and its projection $\tilde{P}$ onto the reference plane.*

Proof: we have that $p' - \tilde{p}' = \alpha w$ where $\alpha = \frac{z - \tilde{z}}{\tilde{z}_o - z_o}$. Since $w = \sigma[T(\tilde{O} - O)]$ and T is a rigid transformation, therefore $\mid w \mid \leq \mid \tilde{z}_o - z_o \mid$. $\square$

Overall scale differences due to translation in depth can be corrected before applying Proposition 6 (see for example [28]), therefore the result applies to similarity transformations as well. The importance of this result is that it suggests that surface shape and motion range can be decoupled, provided that four corresponding points can be identified. The smaller the depth variation between the surface and the reference plane, the larger the range of motion that can be detected from two orthographic views. This can be realized by a two stage computation which starts with a *nominal motion* transformation (first term of equation 8), followed by a *residual motion* computation (the term αw) with the aid of the brightness information. The nominal motion transformation provides a first approximation (determined only by four corresponding points), which is the exact motion for a planar object, leaving a residual whose magnitude is bounded by the depth variation between the surface and the reference plane. The final refinement, determining the residual motion, is provided by the second stage in which the brightness information is used in the form of a second constraint line, as described earlier. This point is developed further below, suggesting a general scheme for measurement of motion.

1.3 Frame of Reference and the Measurement of Motion

The results of section 1.2 suggest that the measurement of motion is conducted relative to a frame of reference, in the form of a reference plane, which determines the direction of motion and the limits on its range (Proposition 6). The range of spatial displacements is bounded by the depth variation between the moving surface and the reference plane. This suggests, therefore, that the frame of reference provides a nominal motion everywhere, which is exact for planar surfaces, by 'pulling' or 'capturing' the motion of all points that are under its influence. The residual motion is later refined by use of local spatio-temporal detectors that implement the constant brightness equation, or any other correlation scheme [32, 52, 1], along the fixed direction determined by the frame of reference.

The notion of a frame of reference that precedes the computation of motion may have some support in human vision literature, although not directly. The phenomenon of 'motion capture' introduced by Ramachandran [38, 39, 40] is suggestive to the kind of motion measurement presented here. Ramachandran and his collaborators observed that the motion of certain salient image features (such as gratings or illusory squares) tend to dominate the perceived motion in the enclosed area by masking incoherent motion signals derived from uncorrelated random dot patterns, in a winner-take-all fashion. Ramachandran therefore suggested that mo-

tion is computed by using salient features that are matched unambiguously and that the visual system assumes that the incoherent signals have moved together with those salient features [38]. The scheme suggested in this paper may be viewed as a refinement of this idea. Motion is 'captured' in Ramachandran's sense for the case of a planar surface in motion, not by assuming the motion of the the salient features but by computing the nominal motion transformation. For a non-planar surface the nominal motion is only a first approximation which is further refined by use of spatio-temporal detectors, provided that the remaining residual displacement is in their range, namely, the surface captured by the frame of reference is sufficiently flat. In this view the effect of capture attenuates with increasing depth of points from the reference plane, and is not affected, in principle, by the proximity of points to the salient features in the image plane.

The motion capture phenomenon also suggests that the salient features that are selected for providing a frame of reference must be spatially arranged to provide sufficient cues that the enclosed pattern is indeed part of the same surface. In other words, not any arrangement of four non-coplanar points, although theoretically sufficient, is an appropriate candidate for a frame of reference. This point has also been raised by Subirana-Vilanova and Richards [44] in addressing perceptual organization issues. They claim that convex image chunks are used as a frame of reference that is imposed in the image prior to constructing an object description for recognition. The frame then determines inside/outside, top/bottom, extraction/contraction and near/far relations that are used for matching image constructs to a model.

Other suggestive data include stereoscopic interpolation experiments by Mitchison and McKee [33]. They describe a stereogram which has a central periodic region bounded by unambiguously matched edges. In certain conditions the edges impose one of the expected discrete matchings (similar to stereoscopic capture, see also [37]). In other conditions a linear interpolation in depth occurred between the edges violating any possible point-to-point match between the periodic regions. The linear interpolation in depth corresponds to a plane passing through the unambiguously matched points, which supports the idea that correspondence starts with the computation of nominal motion, determined by a small number of salient unambiguously matched points, and is later refined using short-range motion mechanisms. Finally, experiments by Todd and Bressan [45] demonstrate that human subjects can determine whether a moving surface is planar from only two orthographic views. This may also suggest that the computation of a frame of reference in the form of planar nominal motion precedes the final computation of motion.

To conclude, the computational results suggest that a long-range mechanism sets up a frame of reference by tracking a selected set of features. The frame pro-

vides a nominal transformation and a matching direction for all other points in the enclosed region. The remaining residual motion following the nominal transformation is handled by short-range motion detectors. This view differs from the classical short-range vs. long-range motion detection in two respects. First it is suggested that the two mechanisms interact in a specific way. Second, the range of detected motion depends not only on the range of the spatio-temporal detectors but also on the three-dimensional shape of the surface, namely, the magnitude of the residual motion depends on how close the enclosed surface is to a plane.

1.4   Implementation

The use of the constant brightness equation for determining the residual motion term αw assumes that $\mid \alpha w \mid$ is small. In practice, the residual motion is not sufficiently small everywhere and, therefore, a hierarchical motion estimation framework is adopted for the implementation. The assumption of small residual motion is relative to the spatial neighborhood and to the temporal delay between frames; it is the ratio of the spatial to the temporal sampling step that is required to be small. Therefore, the smoother the surface the larger the residual motion that can be accommodated. The Laplacian Pyramid [12] is used for hierarchical estimation by refining α at multiple resolutions. The rationale being that large residuals at the resolution of the original image are represented as small residuals at coarser resolutions, therefore satisfying the requirement of small displacement. The α estimates from previous resolutions are used to bring the image pair into closer registration at the next finer resolution.

The particular details of implementation follow the 'warp' motion framework suggested by Bergen and Adelson [9] and by Bergen and Hingorani [10]. Described in a nutshell, a synthesized intermediate image is first created by applying the nominal transformation to the first view. To avoid subpixel coordinates, we actually compute flow from the second view towards the first view. In other words, the intermediate frame at location p contains a bilinear interpolation of the brightness values of the four nearest pixels to the location $\tilde{p}' = A(op) + o' + w$ in the first view, where the 2D affine parameters A, w were computed from view 2 to view 1. The α field is estimated incrementally by projecting previous estimates at a coarse resolution to a finer resolution level. Gaps in the estimation of α, because of vanishing image gradients or other low confidence criteria, are filled-in at each level of resolution by means of membrane interpolation. Once the α field is projected to the finer level, the displacement field is computed (the vector αw) and the two images, the intermediate and the second image, are brought into closer registration. This procedure proceeds incrementally until the finest resolution has been reached.

1.5 Experimental Results

Experiments were done on real imagery of 'Ken', a doll, undergoing rigid rotation, mainly around the vertical axis. Four snapshots were taken covering altogether about 23 degrees of rotation. The light setting consisted of two point light sources located in front of the object, 60 degrees apart from each other.

Three experiments were conducted: (i) long range motion by incrementally adding flow produced by each pair of consecutive images, (ii) long range motion directly, and (iii) establishing approximate correspondence using a single corresponding point and normal flow information.

Privileged points were obtained from flow fields generated by the warp motion algorithm [9, 10] along points having good contrast at high spatial frequencies, e.g. the tip of the eyes, mouth and eye-brows (the location of those points were determined manually).

The combination of the particular light setting and the complexity of the object make it a challenging experiment for the following two reasons: (i) the object is sufficiently complex to have cast shadows and specular points, both of which undergo a different motion than the object itself, and (ii) surface material is dominantly Lambertian and therefore, coupled with the light setting, brightness change will be induced because of change in viewing angle in addition to the change due to motion.

The results of correspondence in all these experiments are displayed in several forms. The flow field is displayed to illustrate the stability of the algorithm, indicated by the smoothness of the flow field. The first image is 'warped' using the flow field to create a synthetic image that should match the second image. The warped image is displayed in order to check for deformations (or lack there of). Finally, the warped image is compared with the second image by superimposing, or taking the difference of, their edge images that were produced using a Canny [15] edge detector with the same parameter settings.

Incremental Long Range Motion

In this experiment, flow was computed independently between each consecutive pair of images, using a fixed set of four privileged points, and then added up to form a flow from the first image, Ken1, to the fourth image, Ken4. The rationale behind this experiment is that because shape is an integral part of computing correspondence/flow, then flow from one consecutive pair to the next should add up in a consistent manner.

Fig. 2 shows the results on the first pair of images, Ken1 and Ken2, separated by 6° rotation. The warped image shows no signs of deformation. As expected, the
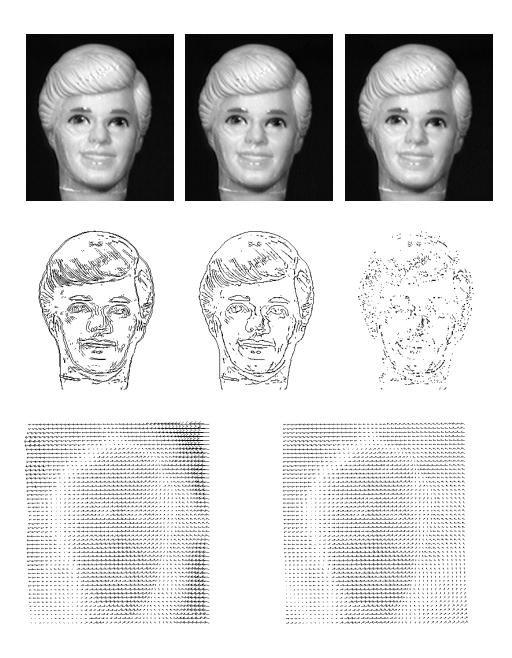
14

Figure 2: Results of shape and correspondence for the pair Ken1 and Ken2. First row: Ken1,Ken2 and the warped image Ken1-2. Second row: edges of Ken1 and Ken2 superimposed, edges of Ken2 and Ken1-2 superimposed, difference between edges of Ken2 and Ken1-2. Third row: flow field in the case where α, the shape constant, is estimated in a least squares manner in a 5×5 sliding window, and flow field when α is computed at a single point (no smoothing).
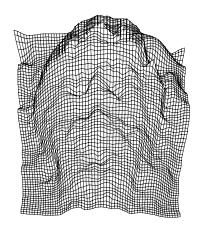
Figure 3: Three-dimensional plot of the shape constant α.

location of strong cast shadows (one near the dividing hair line) and specular points in the warped image do not match those in Ken2. The superimposed edge images illustrate that correspondence is accurate, at least up to a pixel accuracy level. The flow field is smooth even in the case where no explicit smoothing was done. Finally, in Fig. 3 the shape constants α are displayed in a three-dimensional plot. One can clearly see the structure of the head and the bumps and dents corresponding to the location of nose, chin and eyes. One cannot recognize, however, the particular face from this plot or claim that it is a good rendering of a three-dimensional face. The change in brightness due to change in viewing angle is an important cue that is not modeled in this framework, and that may explain the inaccuracies in recovering shape for the images used here. It is also interesting to note the discrepancy between the perceived correspondence, which appears to be accurate, and the true correspondence that would have led to accurate shape constants. This suggests that good correspondence, in the sense of registration, is more attainable than reliable shape descriptors when dealing with real images. More on that, and the relation to visual recognition, in section 2.

Fig. 4 shows the results of adding flow between consecutive pairs computed independently (using the same four privileged points) to produce flow from Ken1 to Ken4. Except the point specularities and the strong shadow at the hair line, the difference between the warped image and Ken4 is only at the level of difference in brightness (because of change in viewing angle). No apparent deformation is observed in the warped image. The flow field is as smooth as the flow from Ken1 to Ken2, implying that the flow was added in a consistent manner.
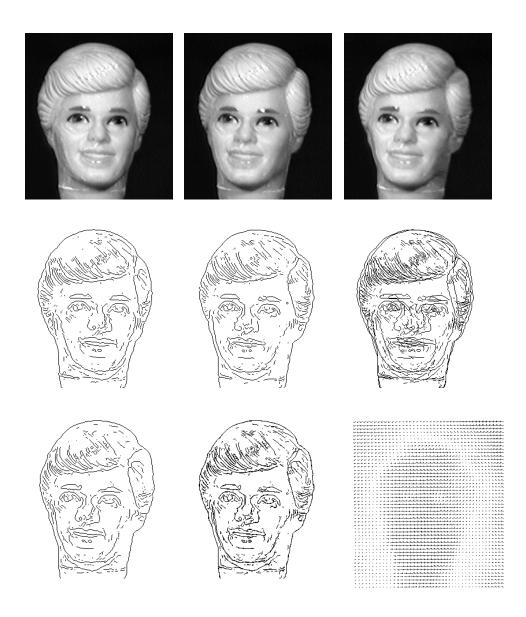
Long Range Motion

Figure 4: Results of adding flow from Ken1 to Ken4. First row: Ken1,Ken4 and the warped image Ken1-4. Second row: edges of Ken1, Ken4 and edges of both superimposed. Third row: edges of Ken1-4, edges of Ken4 and edges of Ken1-4 superimposed, flow field from Ken1 to Ken4 (scaled for display).

The two-stage scheme for measuring motion — nominal motion followed by a short-range residual motion detection — suggests that long-range motion can be handled in an area enclosed by the privileged points. The restriction of short-range motion is replaced by the restriction of limited depth variation from the reference plane. As long as the depth variation is limited, then correspondence should be obtained regardless of the range of motion. Note that this is true as long as we are sufficiently far away from the object's bounding contour. The larger the rotational component of motion — the larger the number of points that go in and out of view. Therefore, we should not expect good correspondence at the boundary. The claim that is tested in the following experiment, is that under long range motion, correspondence is accurate in the region enclosed by the frame of reference, e.g. points that are sufficiently far away from the boundary.

Fig. 5 shows the results of computing flow directly from Ken1 to Ken4. Note the effect of the nominal motion transformation. The nominal motion brings points closer together inside the frame of reference; points near the boundary are taken farther apart from their corresponding points because of the large depth difference between the corresponding object points and the reference plane. The warped image looks very similar to Ken4 except near the boundary of the object. The deformation there may be due to both the relatively large residual displacement, remaining after nominal motion was applied, and to the repetitive intensity structure of the hair; the farther we go from the reference plane the larger the residual displacement αw. Therefore it may be that the frequency of the hair structure caused a misalignment at some level of the pyramid which was propagated.

Approximate Correspondence With a Single Privileged Point

The 2D affine transformation A, w derived from four corresponding points (Proposition 3) describes a constraint line, which together with the constant brightness equation, determines correspondence everywhere else. Also, as shown in appendix 1, any 2D affine transformation that aligns three image points with their corresponding points can be used to define the constraint line, together with one additional corresponding point. It may therefore be possible to look for an affine transformation A, w that approximately aligns 3 points, without actually using 3 corresponding points. Bachelder and Ullman [4] show that measurements of normal flow[5] along at least 6 points determines a 2D affine transformation. Burt *et. al.* [13, 14] show a similar result by deriving a 2D affine transformation directly from the instantaneous brightness measurements using the constant brightness equation.

Following Burt *et. al.* we look for a 2D affine transformation A, w that minimizes

---

[5] the component of $p' - p$ along the image gradient or along the normal to the contour passing through p.
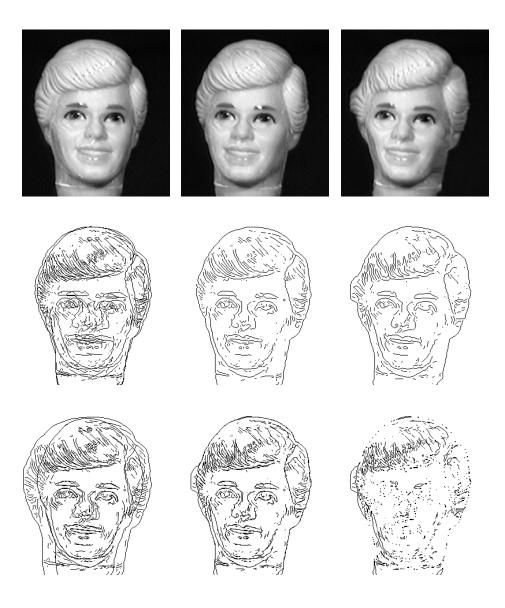
Figure 5: Results of computing long-range flow from Ken1 to Ken4. First row: Ken1,Ken4 and the warped image Ken1-4. Second row: edges of Ken1 and Ken4 superimposed, edges of Ken4 and edges of Ken1-4. Third row: edges of Ken4 superimposed on edges of the nominal transformed Ken1, edges of Ken4 and Ken1-4 superimposed, and difference between edges of ken4 edges of Ken1-4.

the total squared error of the constant brightness equation for which $A(op) + o' + w - p$ is substituted for the unknown velocity. Algebraically, this takes the form:

$$\min_{A,w} \sum_i (\nabla I \cdot (A(op_i) + o' + w - p_i) + I_t)^2.$$

where $o \longleftrightarrow o'$ is a given privileged point. Note that if the area of summation corresponds to a planar patch, then A, w will represent the motion of some plane moving with the object, and therefore accurately aligns at least 3 points with their corresponding points. For a non–planar patch this is not guaranteed, and A, w will only approximately align at least 3 points.

A single region, covering the entire face, was chosen in order to test the accuracy of this scheme on non–planar patches. The affine parameters estimation was performed in an hierarchical framework, and a single privileged point was then chosen (the tip of the left eye). The results of aligning Ken1 and Ken2 are perceptually identical to the four privileged point scheme. The results of aligning Ken1 and Ken3, separated by 14° of rotation, are shown in Fig. 6. Note that although results differ between the four point scheme and the single point scheme, the quality is very similar.

2 Object Recognition and Structure from Motion

The geometrical aspect of visual recognition can be viewed as a problem of compensating for changes in the image induced by changing view positions [50, 26]. Under this view, the visual system must confront similar issues to those dealt with in SFM, albeit in a more implicit manner — one is more concerned in factoring out the effects of shape and viewing transformation on the changing image, rather than recovering them.

Three concepts, that have been recently introduced, seem to play an important role in this view. The first concept is the equivalence between the process of compensating for the change in the image and the process of generating the image from a 2D model [51, 18]. For instance, Ullman and Basri [51] have shown that all possible views that can undergo a similarity transformation in space (rotation, translation and scale), are spanned by the linear combination of three views of the object (two in the case of affine transformation in space, see also result by Poggio [35]). Therefore, any process that can generate a novel view from a 2D model is relevant for purposes of recognition. The second concept, introduced also by Ullman and Basri, is that shape information is equivalent to full correspondence
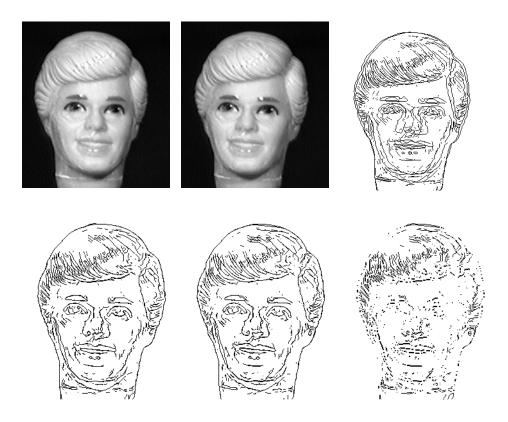
Figure 6: Comparing the four points scheme to the single privileged point scheme.
First row: Ken1,Ken3 and their superimposed edge images. Second row: edges of
Ken3 and Ken1-3 (the warped image) superimposed using four privileged points,
edges of Ken3 and Ken1-3 superimposed using a single privileged point, the differ-
ence between edges of Ken1-3 produced by the four point scheme and the single
point scheme.

among a small set of model views. One therefore does not need to explicitly recover shape and view transformation (motion) in order to generate a novel view — four corresponding points between the novel view and the model views is sufficient for generating the entire view. Finally, the third concept is the distinction between objects with sharp bounding contours and objects with smooth bounding contours [7, 51]. An object with a smooth bounding contour, such as an ellipsoid, does not induce a one-to-one mapping between the object's bounding contours and the projected silhouette. Furtheremore, the bounding contours that generate the silhouette move constantly on the object as the viewing position changes. This case may, therefore, require special attention in generating novel views. Ullman and Basri have shown that for this case the number of views required to approximately span all views undergoing a similarity transformation is five (three in the case of affine transformation in space).

The results derived in section 1 are shown to be relevant to visual recognition in the context of the three concepts described above. In particular, (i) the result that two model views in full correspondence together with four corresponding points with a novel view are sufficient to generate the entire view [51, 35] is rederived using tools from section 1, (ii) a single view can generate novel views taken under similar illumination conditions undergoing limited changes of view position, and (iii) novel views of objects with smooth boundaries can be generated, to a good approximation, from two views in full correspondence.

2.1   Recognition from a Single View

The main result, derived in section 1, is that correspondence can be recovered from two pictures, taken under similar illumination conditions, of an object undergoing an affine transformation in space. The range of allowed viewing transformation was shown to be limited by the structure of the object — the smaller the depth variation, in the region of four corresponding points, the larger the range of viewing transformations. In the context of the first concept, this result is equivalent of saying that a novel view can be generated from a single model view (picture) and four corresponding points, provided the model image and the input image are taken under similar illumination conditions and with a restricted range of viewing transformations.

One straightforward extension is to treat regions of the object as locally flat, and by that to increase the range of viewing transformations for the entire object. This can be implemented by imposing a triangulation on a set of more then four corresponding points [26]. The triangulation divides the image into regions, each with three corresponding points, within which the correspondence method discussed in section 1 can be applied (the fourth corresponding point can be shared among all

triangles).

2.2 Recognition from Two views: Objects with Sharp Boundaries

The basic result, derived by Ullman and Basri [51] and by Poggio [35], is that two model views with full correspondence are sufficient to generate, using the linear combination scheme, a novel view given four corresponding points between the novel view and the model views. Ullman and Basri also pointed out, that with only two model views one cannot distinguish between a non-rigid linear transformation and a rigid transformation of the object.

There are two ways, both straightforward, to re-derive this result in the framework of recovering affine shape. The first derivation follows directly from equations 1 and 2, that for convenience are reproduced below:

$$op = \sum_{j=1}^{3} b_j(op_j)$$

$$o'p' = \sum_{j=1}^{3} b_j(o'p'_j).$$

The affine coordinates can be recovered for every corresponding point, and therefore can be recovered for all points in model view V_1 given full correspondence with model view V_2. Since the affine coordinates are invariant under any affine transformation in space, then given a novel view V and four corresponding points with V_1 and V_2 one can recover the affine coordinates from the known correspondence $V_1 \longleftrightarrow V_2$ and use them to generate V from V_1 (or from V_2). Incidently, this also shows that 1.5 views are sufficient [51, 35] because 2 views provide an over-determined system for solving for the affine coordinates.

One can use a more practical method for generating a novel view by using the constraint line derived in Proposition 3. This can be done in the following way. Let p, p' and p'' be the image coordinates of the point P in the two model views V_1, V_2 and the third novel view V, respectively. Given four corresponding points along the three views one can construct the constraint line, equation 8, between V_1, V_2 and between V_1, V. We take advantage of the separation of shape and motion in equation 8 by noticing that the scale factor α is the same along the constraint line from p to p' and from p to p''. We therefore can find α from the known correspondence p, p' and use that to find the corresponding point in the third view p''. Since α is invariant under affine transformations in space one cannot distinguish between a non-rigid linear transformation and a rigid transformation of the object.

Also, from Propositions 3 and 4, the transformation between the two model views should be other than a pure rotation around the line of sight ($w = 0$ in that case).

2.3   Recognition from Two views: Objects with Smooth Boundaries

In the case of objects with smooth boundaries, the correspondence between the two model views at and near the silhouette no longer relates to the true affine shape parameters at these points. This is because any two corresponding points along the silhouette are projected from different object points.

The fact that the shape parameter α that is recovered from a silhouette point in view V_1 and its corresponding silhouette point in V_2 is not equal to the shape parameter associated with any of the object points projecting to the two corresponding points may work to our advantage. The reason is that the shape parameter α' that is required to correctly generate the same silhouette point in a novel view V also does not relate to a true shape parameter, and therefore it may be expected that $\alpha \approx \alpha'$. It is important to note that as long as the four privileged points are true corresponding points (i.e. not on the silhouette), then the nominal motion transformation and the direction of the constraint line w are correct for all points, including those at and near the bounding contour — it is only the shape parameter α that may be inaccurate at these points.

If indeed α is a good approximation to α', then one can use the same method for generating novel views as that used for objects with sharp boundaries — with the same number of model views.

The following section analyzes the accuracy of this method under the assumption of pure rotation around y axis (rotation around z axis can be neglected), reference plane ortho-parallel, and that rim points are on locally spherical patches. Under these assumptions, the error relative to the radius of curvature at the rim is shown to be typically less than 3% for relatively large rotations (30 degrees) and less than 1% for a 15 degree rotation. Experimental results follow.

2.4   Analysis of the Prediction Method

For the purpose of analysis, one can ignore rotation around the z axis, translation, and scaling. I further make the following simplifications: (i) reference plane is ortho-parallel to the image plane, (ii) rotation is only around the y axis, and (iii) the boundary points projecting to the silhouettes are on locally spherical patches, with radius r.
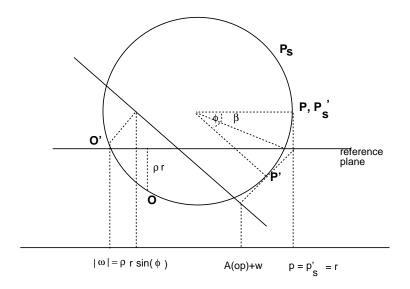
Figure 7: Cross section of a sphere perpendicular to the vertical axis. See text for reference.

Fig. 7 shows a cross section of a sphere, that is perpendicular to the y axis, and a point P on its rim. The point P'_s is the new rim point followed by a ϕ degree rotation around the y axis. Let the reference plane be at a distance $r \sin \beta$ from the center of the sphere, and let the privileged point O be located on the sphere such that the distance $z_o - \tilde{z}_o = r\rho$, for some constant ρ. We therefore have that $|w| = r\rho \sin \phi$. The nominal transformation associated with P is the projection of $\tilde{P}'$ which is equal to $|A(op) + w| = r(\cos \phi - \sin\phi \sin \beta)$. The shape parameter α scales w to satisfy the equation (only the x-component is displayed): $A(op) + w + \alpha w = r$, and therefore

$$\alpha = \frac{1 - \cos \phi + \sin \phi \sin \beta}{\rho \sin \phi}.$$

We use α to predict the new location of the corresponding point p' resulting from some other angle of rotation $\hat{\phi}$ around the Y axis. The motion component and the length of w corresponding to the new angle $\hat{\phi}$ are $r(\cos \hat{\phi} - \sin\hat{\phi} \sin \beta)$ and $r\rho \sin \hat{\phi}$, respectively. Noting that an exact correspondence will set $p' = r$ for any angle $\hat{\phi}$, the error, relative to the radius r, is therefore:

$$\epsilon = \cos \hat{\phi} + \frac{(1 - \cos \phi) \sin \hat{\phi}}{\sin \phi}.$$
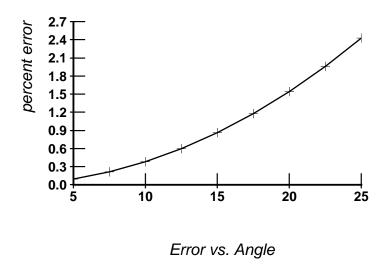
25

Error vs. Angle

Figure 8: Percentage of relative error as a function of the angle between the model views, assuming worst case interpolation error.

The worst case error for interpolation, i.e., $\hat{\phi} < \phi$, is when $\hat{\phi} = \arctan\left(\frac{1-\cos\phi}{\sin\phi}\right) = \frac{\phi}{2}$. Fig. 8 shows the percentage of error as a function of ϕ, taking $\hat{\phi}$ to be the worst case interpolation error. We see that the more distant the two model views — the larger the relative error. Also, the absolute error increases with the radius r, for example, the lower the curvature along the line of sight the larger the absolute error. The expected worst case absolute error for generating views of 'Ken', given Ken1 and Ken4 as the model views, are 1.5 pixels for Ken2 and 2 pixels for Ken3. This is because the projected radius is about 100 pixels, $\phi = 23$ and $\hat{\phi} = 6, 9$ for Ken2 and Ken3, respectively.

Experimental results shown below confirm these estimates. Full correspondence were obtained using the incremental flow estimation described earlier (results that were shown in Fig. 4). Four corresponding points were manually chosen among the two model views and between Ken2 and Ken3. Fig. 9 shows the results of generating Ken2 and Ken3. As expected, the errors in the silhouette of Ken2 are smaller than those in Ken3. This is because Ken3 is further apart from the model views than Ken2, as illustrated in the figure. The errors along the silhouette of Ken2 are less than 1 pixel for most of the points, and along other silhouette points the error is between 1 to 2 pixels. The errors along the silhouette of Ken3 are between 1 and 2 pixels.

In conclusion, a method for generating novel views from two model views was suggested. The method is based on the principle of shape and motion separation
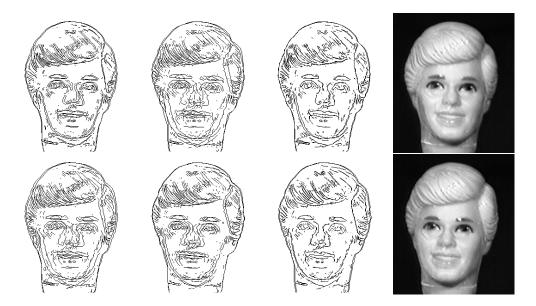
Figure 9: Generating novel views from two model views, Ken1 and Ken4. The top row shows results of generating Ken2 from Ken1, and the bottom row shows the results of generating Ken3 from Ken4. The first two images illustrate the distance between the novel view and the two model views, the third image is an overlay of the edges of the original view and the predicted view, the fourth image is the generated view.

derived in Proposition 3. The method is accurate for objects with sharp bounding contours, and can handle, to a good approximation, objects with smooth bounding contours. An analytic analysis followed by experimental results on images of 'Ken' illustrate the accuracy of the method.

In comparing the affine-shape recognition scheme to the linear combination scheme of Ullman and Basri, one sees the following tradeoff: the linear combination scheme is theoretically more accurate in the case of objects with smooth boundaries (see Basri [5], for analysis) than the method suggested here. This is at the expense of having more views in the model, and having to include points along the bounding contour in the sample of privileged points (otherwise the linear coefficients are not unique). The method suggested here does not require privileged points along the boundary, which makes it easier to find a small number of reliable points to use for recognition.

3   Summary

The paper presented a model for recovering affine shape and correspondence from two orthographic views for the purposes of structure from motion and object recognition. It was shown that it is possible to recover shape and full correspondence/flow simultaneously, by using the instantaneous change in brightness, together with four corresponding points, as an integral part of the computational model. It was shown that a 2D affine transformation, derived directly from the four corresponding points, represents a constraint line that captures both the affine shape and the motion of a plane, that serves as a frame of reference, imposed on the object (Proposition 3). Based on that result, it was suggested that the measurement of motion starts by imposing a frame of reference that is defined by a small number of salient, unambiguously matched, features in the image. The motion of the frame 'captures' the motion of the remaining image points and takes them part of the way towards their corresponding points. Motion is then refined using local spatio-temporal detectors that are tuned along a known direction in the image. The magnitude of the refinement is bounded by the depth variation between the surface and the frame (Proposition 6).

Those results were shown to apply to visual recognition by generating novel views of sharp and smooth boundary objects from two model views, or from a single view but with a restricted viewing transformation range.

Appendix 1: Alternative Algebraic Form of the Constraint Line

Koenderink and Van-Doorn [28] show that the constraint line can be derived in two stages, first p is represented in the 2D affine frame defined by 3 of the corresponding points, and then the fourth point is used to find the third affine vector by subtracting its corresponding point from its projection in the 2D affine frame.

This result is derived below in a single step algebraic proof which also shows that the resulting constraint line is a particular case of equation 5 in which $\tilde{r}_3 = 0$. Also shown is the result that a 2D affine transformation aligning at least 3 points with their corresponding points can be used, together with an additional corresponding point, to derive a constraint line of the type introduced by Koenderink and Van-Doorn.

Let A and w (not the same A, w as in Proposition 3) be the 2D affine transformation that align 3 of the corresponding points o, p_1, p_2, i.e. $o' = Ao + w$ and $p'_j = Ap_j + w$, $j = 1, 2$. By subtracting the first equation from the other two we get:

$$o'p'_1 = A(op_1)$$
$$o'p'_2 = A(op_2)$$

and therefore A is the matrix $[o'p'_1, o'p'_2][op_1, op_2]^{-1}$. Let p_3 be the fourth privileged point, and let p be an arbitrary point. We therefore have:

$$o'p' = \sum_{j=1}^{3} b_j(o'p'_j) = b_1 A(op_1) + b_2 A(op_2) + o'p'_3 + b_3 A(op_3) - b_3 A(op_3)$$
$$= A(op) + b_3(o'p'_3 - A(op_3))$$

and considering that $w = o' - Bo$ we get:

$$p' = Ap + w + b_3(p'_3 - Ap_3 - w)$$

In this case the third frame vector is not in direction of w, but as Koenderink and Van-Doorn noted is the result of subtracting the projected motion of P_3 onto the reference plane passing through O, P_1, P_2 from p'_3.

Appendix 2: Proof of Proposition 4

Proposition 4 *The 2D affine transformation defined in Proposition 3 admits the following interpretation: the affine vector w is the projection of the vector $\tilde{O}' - O'$ onto the image plane and is perpendicular to the xy projection of the rotation axis of the transformation T. If T is a similarity transformation (rotation, translation and scale), then α associated with the point p is equal to:*

$$\alpha = \frac{z - \tilde{z}}{\tilde{z}_o - z_o}$$

where $z, \tilde{z}, z_o$ and $\tilde{z}_o$ are the depth values of $P, \tilde{P}, O$ and $\tilde{O}$, respectively.

Proof: Geometrically, any point P that is coplanar with P_1, P_2, P_3 can be represented as a convex combination of the three vectors OP_j, therefore α of the point p is equal to 0. This in particular shows, in a very simple manner, that a 2D affine transformation accounts for the projected motion of a plane [42]. This also proves that α represents the deviation of P, along the line of sight, from the reference plane.

The vector OP can be represented as sum of the following two vectors:

$$OP = P - O = (P - \tilde{P}) + (\tilde{P} - O).$$

Applying an affine transformation T followed by an orthographic projection yields:

$$o'p' = \sigma[T(P - \tilde{P})] + \sigma[T(O\tilde{P})].$$

Since $\tilde{P}$ is on the reference plane, we have that $\sigma[T(O\tilde{P})] = A(op) + w$. We therefore have the following:

$$\sigma[T(P - \tilde{P})] = \alpha w$$

where $\alpha = \sum b_j - 1$. Using the same reasoning, we get that

$$w = \sigma[T(\tilde{O} - O)].$$

We therefore see that w is the projection of $\tilde{O} - O$ at the second time cinstance. Because $\tilde{O} - O$ is parallel to $P - \tilde{P}$, we may represent the deviation of P from the reference plane as a scale factor of $\tilde{O} - O$. Furthermore, since $\tilde{O} - O$ is along the line of sight, then the projection $\sigma[T(\tilde{O} - O)]$ is perpendicular to the xy projection of the rotation axis.

If T is a rigid transformation, possibly followed by uniform scaling, then the relationship between $P - \tilde{P}$ and $\tilde{O} - O$ remains fixed, before and after T is applied, and therefore α, the scale factor becomes:

$$\alpha = \frac{z - \tilde{z}}{\tilde{z}_o - z_o}$$

where $z, \tilde{z}, z_o$ and $\tilde{z}_o$ are the depth values of $P, \tilde{P}, O$ and $\tilde{O}$, respectively. □

31

References

[1] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America*, 2:284–299, 1985.

[2] G. Adiv. Inherent ambiguities in recovering 3D motion and structure from a noisy flow field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11(5):477–489, 1989.

[3] P. Anandan. A unified perspective on computational techniques for the measurement of visual motion. In *Proceedings Image Understanding Workshop*, pages 219–230, Los Angeles, CA, February 1987. Morgan Kaufmann, San Mateo, CA.

[4] I.A. Bachelder and S. Ullman. Contour matching using local affine transformations. In *Proceedings Image Understanding Workshop*. Morgan Kaufmann, San Mateo, CA, 1991. To Appear.

[5] R. Basri. *The recognition of 3-D solid objects from 2-D images*. PhD thesis, Weizmann Institute of Science, Rehovot, Israel, 1990.

[6] R. Basri. On the uniqueness of correspondence under orthographic and perspective projections. In *Proceedings Image Understanding Workshop*. Morgan Kaufmann, San Mateo, CA, 1991. To Appear.

[7] R. Basri and S. Ullman. The alignment of objects with smooth surfaces. In *Proceedings of the International Conference on Computer Vision*, pages 482–488, Tampa, FL, December 1988.

[8] B.M. Bennet, D.D. Hoffman, J.E. Nicola, and C. Prakash. Structure from two orthographic views of rigid motion. *Journal of the Optical Society of America*, 6:1052–1069, 1989.

[9] J.R. Bergen and E.H. Adelson. Hierarchical, computationally efficient motion estimation algorithm. *Journal of the Optical Society of America*, 4:35, 1987.

[10] J.R. Bergen and R. Hingorani. Hierarchical motion-based frame rate conversion. Technical report, David Sarnoff Research Center, 1990.

[11] T. Broida, S. Chandrashekhar, and R. Chellapa. recursive 3-d motion estimation from a monocular image sequence. *IEEE Transactions on Aerospace and Electronic Systems*, 26:639–656, 1990.

[12] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communication*, 31:532–540, 1983.

[13] P.J. Burt, J.R. Bergen, R. Hingorani, R. Kolczinski, W.A. Lee, A. Leung, J. Lubin, and H. Shvaytser. Object tracking with a moving camera, an application of dynamic motion analysis. In *IEEE Workshop on Visual Motion*, pages 2–12, Irvine, CA, March 1989.

[14] P.J. Burt, J.R. Bergen, R. Hingorani, S. Peleg, and P. Anandan. Dynamic analysis of image motion for vehicle guidance. In *IEEE Workshop on Intelligent Motion Control*, pages 75–82, Bogazici University, Istanbul, August 1990.

[15] J. Canny. Finding edges and lines in images. A.I. TR No. 720, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1983.

[16] C.H. Chien and J.K. Aggarwal. Shape recognition from single silhouette. In *Proceedings of the International Conference on Computer Vision*, pages 481–490, London, December 1987.

[17] R. Dutta and M.A. Synder. Robustness of correspondence based structure from motion. In *Proceedings of the International Conference on Computer Vision*, pages 106–110, Osaka, Japan, December 1990.

[18] S. Edelman and T. Poggio. Bringing the grandmother back into the picture: a memory based view of object recognition. A.I. Memo No. 1181, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[19] O.D. Faugeras and M. Hebert. The representation, recognition and location of 3D objects. *International Journal of Robotics Research*, 5(3):27–52, 1986.

[20] M.A. Fischler and R.C. Bolles. Random sample concensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–395, 1981.

[21] E.C. Hildreth. Computations underlying the measurement of visual motion. *Artificial Intelligence*, 23(3):309–354, August 1984.

[22] B.K.P. Horn. *Robot Vision*. MIT Press, Cambridge, Mass., 1986.

[23] B.K.P. Horn and B.G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[24] B.K.P. Horn and E.J. Weldon. Direct methods for recovering motion. *International Journal of Computer Vision*, 2:51–76, 1988.

[25] T.S. Huang and C.H. Lee. Motion and structure from orthographic projections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11:536–540, 1989.

33

[26] D.P. Huttenlocher and S. Ullman. Object recognition using alignment. In *Proceedings of the International Conference on Computer Vision*, pages 102–111, London, December 1987.

[27] J.J. Koenderink. *Solid Shape*. MIT Press, Cambridge, MA, 1990.

[28] J.J. Koenderink and A.J. van Doorn. Affine structure from motion. *Journal of the Optical Society of America*, 8:377–385, 1991.

[29] Y. Lamdan and H.J. Wolfson. Geometric hashing: A general and efficient model–based recognition scheme. In *Proceedings of the International Conference on Computer Vision*, pages 238–249, Osaka, Japan, Dec. 1990.

[30] H.C. Longuet-Higgins and K. Prazdny. The interpretation of a moving retinal image. *Proceedings of the Royal Society of London B*, 208:385–397, 1980.

[31] D. Marr and T. Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London B*, 204:301–328, 1979.

[32] D. Marr and S. Ullman. Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London B*, 211:151–180, 1981.

[33] G.J. Mitchison and S.P. McKee. interpolation in stereoscopic matching. *Nature*, 315:402–404, 1985.

[34] A. Pentland. Photomertric motion. In *Proceedings of the International Conference on Computer Vision*, pages 178–187, Osaka, Japan, December 1990.

[35] T. Poggio. 3D object recognition: on a result of Basri and Ullman. Technical Report IRST 9005-03, May 1990.

[36] K. Prazdny. Egomotion and relative depth map from optical flow. *Biological Cybernetics*, 36:87–102, 1980.

[37] K. Prazdny. 'capture' of stereopsis by illusory contours. *Nature*, 324:393, 1986.

[38] V.S. Ramachandran. Capture of stereopsis and apparent motion by illusory contours. *Perception and Psychophysics*, 39:361–373, 1986.

[39] V.S. Ramachandran and P. Cavanagh. subjective contours capture stereopsis. *Nature*, 317:527–530, 1985.

[40] V.S. Ramachandran and V. Inada. spatial phase and frequency in motion capture of random-dot patterns. *Spatial Vision*, 1:57–67, 1985.

[41] J.W. Roach and J.K. Aggarwal. Computer tracking of objects moving in space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:127–135, 1979.

[42] J.G. Semple and G.T. Kneebone. *Algebraic Projective Geometry*. Clarendon Press, Oxford, 1952.

[43] A. Shashua. Illumination and view position in 3D visual recognition. In S.J. Hanson J.E. Moody and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*. San Mateo, CA: Morgan Kaufmann Publishers, 1992.

[44] J.B. Subirana-Vilanova and W. Richards. perceptual organization, figure-ground, attention and saliency. A.I. Memo No. 1218, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, August 1991.

[45] J.T. Todd and P. Bressan. The perception of 3D affine structure from minimal apparent motion sequences. *Perception and Psychophysics*, 48:419–430, 1990.

[46] C. Tomasi. *shape and motion from image streams: a factorization method*. PhD thesis, School of Computer Science, Carnagie Mellon University, Pittsburgh, PA, 1991.

[47] C. Tomasi and T. Kanade. Factoring image sequences into shape and motion. In *IEEE Workshop on Visual Motion*, pages 21–29, Princeton, NJ, September 1991.

[48] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, Cambridge and London, 1979.

[49] S. Ullman. Computational studies in the interpretation of structure and motion: summary and extension. In J. Beck, B. Hope, and Azriel Rosenfeld, editors, *Human and Machine Vision*. Academic Press, New York, 1983.

[50] S. Ullman. Aligning pictorial descriptions: an approach to object recognition. *Cognition*, 32:193–254, 1989. Also: in MIT AI Memo 931, Dec. 1986.

[51] S. Ullman and R. Basri. Recognition by linear combination of models. A.I. Memo No. 1052, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, August 1989.

[52] J.P.H. Van Santen and G. Sperling. Elaborated reichardt detectors. *Journal of the Optical Society of America*, 2:300–321, 1985.

[53] A. Verri and T. Poggio. Motion field and optical flow: Qualitative properties. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:490–498, 1989.

[54] A.M. Waxman and S. Ullman. Surface structure and 3-D motion from image flow: a kinematic analysis. *International Journal of Robotics Research*, 4:72–94, 1985.

[55] A.M. Waxman and K. Wohn. Contour evolution, neighborhood deformation and global image flow: Planar surfaces in motion. *International Journal of Robotics Research*, 4(3):95–108, Fall 1985.