# 3D object recognition and matching: on a result of Basri and Ullman*

Tomaso Poggio

The main point of this note is to characterize the algebraic structure of the views of one 3D object under orthographic projection. Consider the linear vector space $\mathcal{R}^{3N}$ of 3D views of all objects, with a 3D view being the vector of the $x$, $y$ and $z$ coordinates of each of $N$ feature points. Consider the subspace $V^{3N}_{ob_i}$ generated by one view of a specific object and by the action on it of the group of *uniform* linear transformations $\mathcal{L}$ (i.e. the same linear transformation is applied to each feature point). $\mathcal{L}$ is an algebra of order 9, and therefore a linear vector space isomorphic to $\mathcal{M}_3$ (that is the space of the $3 \times 3$ matrices with real elements). Thus, $V^{3N}_{ob_i}$ is a linear vector space isomorphic to $\mathcal{R}^9$. The projection operator (orthographic projection) that deletes the $z$ components from the 3D views, maps $V^{3N}_{ob_i}$ into a linear vector subspace $V^{2N}_{ob_i}$, isomorphic to $\mathcal{R}^6$. $V^{2N}_{ob_i}$ consist of vector with $x$ and $y$ components and can be written as the direct sum $V^{2N}_{ob_i} = V^N_x \oplus V^N_y$, where $V^N_x$ and $V^N_y$ are non-intersecting linear subspaces, each isomorphic to $\mathcal{R}^3$. In addition, I have proved (Basri has obtained this result independently) that $V^N_x = V^N_y$, which implies that 1.5 snapshots are sufficient for "learning" an object (generically). If 3D translations are included, a linear subspace, isomorphic to $\mathcal{R}^2$ must be added to the linear space spanned by the 2D views of one object. The 1.5 views theorem implies that the $x$ and the $y$ vectors obtained from the 2 frames are linearly dependent. This in turn implies that 4 matched points across two views are sufficient (generically) to determine 1-D epipolar lines for matching all other points. This is a very useful result (first obtained in a different context by Huang and Lee (1989)) in correspondence problems involving 2 frames and affine, uniform transformations in $3D$.

## 1. Introduction

Basri and Ullman (1990) have recently discovered the striking fact that under orthographic projection a view of a 3D object is the linear combination of a small number of views of the same object. In this note, we reformulate

* Most of the content of this paper has appeared as IRST Technical Report 9005–03, 1990.

their results in the more abstract setting of linear algebra. This framework makes the result very transparent: the constraint of linear transformation (the same linear transformation for each vertex) implies immediately that the set of views of an object spans a 9-dimensional space, independently of the number of vertices; orthographic projection preserves linearity while reducing the number of dimensions to 6. Simple considerations show that the linear spaces of the $x$ and $y$ coordinates are nonintersecting and that each has dimension 3. Furthermore I prove that they are equivalent, implying that 1.5 snapshots are sufficient to learn the model of one object.

## 2. Any view of a 3D object is a linear combination of a small, fixed number of views

This section provides the main result of Basri and Ullman (in the second subsection).

*2.1. Any 3D-view of an object is a linear combination of 9 views*

Let us define a *3D-view* of a 3D object as:

$$\mathbf{X}^{obj} = \begin{pmatrix} x_1 \\ y_1 \\ z_1 \\ x_2 \\ y_2 \\ z_2 \\ . \\ . \\ . \\ x_n \\ y_n \\ z_n \end{pmatrix}$$

with $X \in \mathcal{R}^{3n}$, which is a vector space in the usual way.

I consider the set of *uniform* (my definition) linear operators on $\mathcal{R}^{3n}$, defined by the $3n \times 3n$ matrices $\mathbf{L}^{3n}$, where $\mathbf{L}^{3n} = \mathbf{I}_n \otimes L$ is the tensor product of $\mathbf{I}_n$ and $L$:

$$\mathbf{L}^{3n} = \begin{pmatrix} L & 0 & . & 0 \\ 0 & L & . & 0 \\ . & . & . & . \\ 0 & 0 & . & L \end{pmatrix}$$

where

$$L = \begin{pmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{pmatrix}$$

is an affine transformation on $\mathcal{R}^3$. Translation in $3D$ space is taken care of separately (see later).

The space of the $\mathbf{L}^{3n}$ operators is a vector space which is *isomorphic* to the vector space of the $L$ matrices. It therefore has a basis of 9 elements independently of $n$.

I can express

$$\mathbf{L}^{3n} = \sum_{i=1}^{9} a_i \mathbf{L}_i^{3n}$$

where $a_i$ can be identified with the appropriate $l_{i,j}$ and $L_i^{3n}$ with the usual basis for $L^{3n}$, i.e. with the elementary matrices $E$,
and thus

$$\mathbf{X} = \mathbf{L}^{3n}\mathbf{X}_0 = \sum_{i=1}^{9} a_i \mathbf{L}_i^{3n}\mathbf{X}_0 = \sum_{i=1}^{9} a_i \mathbf{X}_i$$

where $\mathbf{X}_i$ are 9 independent 3D views of the specific object, needed to span the 9 elements of $L$, 3 for each coordinate, and $X_0$ is a particular view chosen as the "initial" view. Thus:

**Theorem 2.1** *The vector space $V_{0b}^{3D}$ generated by the action of uniform linear transformations on a 3D view of a specific object is a 9-dimensional subspace of $\mathcal{R}^{3n}$, 3 dimensions for $x$, 3 for $y$ and 3 for $z$.*

Thus any object $ob_i$ generates a corresponding low dimensional subspace $V_{ob_i}^{3D}$ of all possible views of all objects ($\mathcal{R}^{3n}$). Of course, $V_{ob_i}^{3D} \not\equiv \mathcal{R}^{3n}$, *iff* $n > 3$. In other words, to have object specificity, i.e., for this result to be *nontrivial*, it is necessary that $n > 3$. Notice that $\mathcal{R}^{3n} = V_{ob_i} + V_{ob_2} + \ldots$.

*2.2. Any 2D-view of a 3D object is a linear combination of 6 2D-views*

Now consider the orthographic projection $P : \mathcal{R}^{3n} \to \mathcal{R}^{2n}$, defined by $P\mathbf{X} = \mathbf{x}$, that is

$$P \begin{pmatrix} x_1 \\ y_1 \\ z_1 \\ x_2 \\ y_2 \\ z_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \\ y_n \\ z_n \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \\ x_3 \\ y_3 \\ \cdot \\ \cdot \\ \cdot \\ x_n \\ y_n \end{pmatrix}$$

with $P$ being a linear operator with the matrix representation

$$P = \begin{pmatrix} 1 & 0 & . & . & . & . & . & . & 0 \\ 0 & 1 & 0 & . & . & . & . & . & 0 \\ 0 & 0 & 0 & 1 & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . & . & . \\ . & . & . & 0 & 0 & . & 1 & 0 & 0 \\ 0 & 0 & . & . & . & . & 0 & 1 & 0 \end{pmatrix}$$

We define $\mathbf{x}$ as the *2D-view* of a 3D object.

The result below follows immediately (6 views span the elements of $L$ in the first 2 rows) and is the main result of Basri and Ullman (in a different formulation):

**Theorem 2.2** *The vector space $V_{ob_i}$ given by $V_{ob_i} = PV_{ob_i}^{3D}$ is a six-dimensional subspace of $\mathcal{R}^{2n}$ (the space of all 2D orthographic views of all 3D objects), i.e. $\mathbf{x_{ob}} = \sum_{i=1}^{6} a_i \mathbf{x}_{ob}^i$.*

**Remark:** The inclusion of rigid translations is equivalent to the addition of a two-dimensional linear subspace (the same for all objects), spanned by the vectors

$$\mathbf{t_x} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ . \\ . \\ . \end{pmatrix}$$

and

$$\mathbf{t_y} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ . \\ . \\ . \end{pmatrix}$$

## 3. The $x$ and the $y$ coordinates of a view are each a separate linear combination of 3 views

In the previous section we have seen that any *2D-view* of a 3D object under orthographic projection is the linear combination of 6 *2D-views*. This section reformulates another observation of Basri and Ullman: the $x$ coordinates of a *2D-view* are a linear combination of the $x$ coordinates of 3 *2D-views* and the $y$ coordinates are a linear combination of the $y$ coordinates of 3 *2D-views*, the two combinations being independent of each other.

Let us consider a similarity transformation of $\mathbf{x}$:

$$T\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ \cdot \\ x_n \\ y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \\ z_1 \\ \cdot \\ \cdot \\ \cdot \\ z_n \end{pmatrix}$$

Under this similarity transformation, $\mathbf{L}^{3n}$ becomes a $3\times3$ matrix of 9 (that is $3\times3$) blocks. Each block is a multiple of $I \in \mathcal{R}^{n,n}$ (notice the "isomorphism" to $L$!).

$$T^T LT = \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix}$$

where

$$I_{11} = \begin{pmatrix} l_{11} & 0 & 0 & . & . & . \\ 0 & l_{11} & 0 & . & . & . \\ 0 & 0 & l_{11} & . & . & . \\ . & . & . & . & . & . \end{pmatrix}$$

and so on for the other blocks.
The same argument of the previous section makes it clear that defining

$$\xi = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}$$

$$\eta = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}$$

the following holds:

$$\xi = \sum_{i=1}^{3} l_{1i}\xi_i$$

$$\eta = \sum_{i=1}^{3} l_{2i}\eta_i.$$

Thus we have proved:

**Theorem 3.1** *The subspace spanned by the vectors $\xi$ — the $x$ components of $\mathbf{x}$ — which is a $n-$dimensional subspace of $V_{ob}^{2D}$ (which is 2n-dimensional), is spanned by three views of the $x$ coordinates of the object undergoing uniform transformations, i.e., each $\xi$ can be represented as the linear combination of 3 independent $\xi_i$. The same is true for the $\eta$: each $\eta$ is an independent linear combination of 3 independent $\eta_i$. Again, $n > 3$ in order for this to be non-trivial (since $\xi \equiv \mathcal{R}^n$ for $n \leq 3$).*

**Remark:** The basis of $\xi$ and the basis of $\eta$ depend on the specific object.

## 4. $V_x$ and $V_y$ have the same basis, i.e. 1.5 snapshots suffice

We know from the previous sections that $V_{ob_i}^{2N} = V_x^N \oplus V_y^N$, where $dim V_x = dim V_y = 3$. A stronger property holds:

**Theorem 4.1** *$V_x = V_y$ (in the sense of Furlanello)*

*Proof.* Assume that $V_x$ and $V_y$ are not identical (I consider the projections of the $x$ and $y$ components expressed originally in the same base in $V$) restrictions: then there is a vector $\mathbf{y}$ which is in $V_y$ and not in $V_x$ (or viceversa). Then I can take the 3D view that originated $\mathbf{y}$ (through orthogonal projection) and apply to it a legal transformation consisting of a rigid rotation of 90 degrees in the image plane (such a transformation is in $L$ and therefore is legal). The $x$ view of that 3D vector is the $\mathbf{y}$, contradicting the assumption. It follows that $V_x = V_y$.

**Remarks**

1 The same argument shows that $V_x = V_y = V_z$

2 The same basis of three vectors spans $V_x$ and $V_y$ (separately).

3 The property that the $x$ views and the $y$ views of the same 3D object from the same snapshot are independent is generic, since if they were dependent, a very slightly different object, differing only in the $y$ coordinate of one vertex would have independent views (Bruno Caprile, pers. com.).

4 In general, 1.5 snapshots are sufficient to provide a basis

5 Any 4 vectors from $V_x$ and $V_y$ are linearly dependent.

## 5. A corollary of the 1.5 views theorem: given four matched points, correspondence for motion or recognition is easy

A direct consequence of the above 1.5 views theorem is that the 4 vectors (from 2 orthographic views) of the **x** and **y** components of an object undergoing an uniform affine transformation in 3D (in particular a rigid transformation in 3D) are linearly dependent, that is

$$\alpha_1 \mathbf{x}_1 + \beta_1 \mathbf{y}_1 + \alpha_2 \mathbf{x}_2 + \beta_2 \mathbf{y}_2 = 0$$

This implies that the correspondence of at least 4 points (including translations) in two frames determines epipolar lines for the matching of all other points (the observation is due to Ronen Basri; see also Amnon Sha'shua; a similar result — but not this proof — was first obtained by Lee and Huang, 1988). This means that for each point $(x_1, y_1)$ in frame 1 the corresponding point in frame 2 satisfies the equation

$$y = mx + A$$

with $m = \alpha_2^*$ and $A = -(\alpha_1^* x_1 + \beta_1^* y_1)$ and $\alpha_1^* = \alpha_1/\beta_2$ and so on. Translations are taken care of by matching one point in the two frames. Three additional "generic" points are needed to solve for $\alpha_1^*$, $\alpha_2^*$ and $\beta_1^*$.

Therefore in problems of matching between 2 frames — in motion or recognition — four points are sufficient to determine epipolar lines along which the matching of the other points can be more easily found.

## 6. The case of rigid transformations, i.e., rotations in 3D

The previous two section have considered the case of *uniform linear transformations* in 3D of a 3D object. The space of such transformations is a vector space that contains as a nonlinear subspace the group of the rigid rotations in 3D (which is easily seen not to be a vector space). Can we characterize what the restriction to rigid rotations means? This section addresses this question.

Consider the restriction $L = R$ with $R^T R = I$. Then:

$$\begin{cases} l_{11}^2 + l_{12}^2 + l_{13}^2 = 1 \\ l_{11} l_{21} + l_{12} l_{22} + l_{13} l_{23} = 0 \\ l_{21}^2 + l_{22}^2 + l_{23}^2 = 1 \end{cases}$$

The equations define a nonlinear subspace of the space $\xi = \{l_{11}, l_{12}, l_{13}\}$ isomorphic to $\mathcal{R}^3$, and of $\eta = \{l_{21}, l_{22}, l_{23}\}$, also isomorphic to $\mathcal{R}^3$. Of course, $\xi$ is a linear subspace of $\mathcal{R}^n$, the *space of all views of the x coordinates of all*

*objects*. Rotations are the intersection of $\xi$ with the conics defined by the previous equations.

The 2D views of one object defined by uniform affine transformations span $\{l_{11}, l_{12}, l_{13}\} = \mathcal{R}^3$. The 2D views of one object defined by rigid transformations, i.e., rotations, span a nonlinear subspace of $\mathcal{R}^3$, namely, the surface of the unit sphere in $\mathcal{R}^3$. All points on the *unit* sphere are allowed for $\{l_{11}, l_{12}, l_{13}\}$ (thus we "use up" two parameters). The triplet $(l_{21} l_{22} l_{23})$ is determined as one parameter family. Geometrically, once the vector $l_{11}, l_{12}, l_{13}$ is fixed on the unit sphere, an orthogonal circle is determined on which the vector $(l_{21}, l_{22}, l_{23})$ must lie.

## Acknowledgements

Daphna Weinshall explained to me that Basri and Ullman's result was not restricted to rigid rotations despite what I had read in their original Weizman Technical Report. Bruno Caprile and Federico Girosi provided several useful suggestions. Shimon Ullman told me about the futility of trying to extend the last section because of Huang's recent results on structure from motion. Cesare Furlanello was the only person to read it carefully and to explain to me what the little mathematics in it really means.