Multiple-view Geometry and Photometry

Amnon Shashua
Technion — Israel Institute of Technology
Department of Computer Science
Haifa 32000, Israel.
e-mail: shashua@cs.technion.ac.il

Abstract

The issue of how to represent and manipulate the information arising from a multitude of perspective pictures of a 3D scene is a recent and growing topic of interest. In this paper we present a summary account of up-to-date research, my own and with colleagues, on this topic, including: the trilinear constraints and associated tensor; the properties of the trilinear tensor and their relevance to camera geometry and invariance; rank deficiencies and $N > 4$ multiple-view geometry; and the photometric duality in the form of the "Tensor Brightness Constraint".

1 Introduction

The algebraic and geometric relations across multiple perspective views is a recent and growing interest which is relevant to a number of topics including (i) issues of 3D reconstruction from 2D data, (ii) representations of visual scenes from video data, (iii) image synthesis and animation, and (iv) visual recognition and indexing.

Typical to these topics is the question about the limitations and possibilities of going from two-dimensional (2D) measurements of point matches (correspondences) across two or more views to properties of the three-dimensional (3D) object or scene. Since the relationship between the 3D world and the 2D image space combines together 3D shape parameters, camera viewing parameters and 2D image measurements, the question of limitations and possibilities, in its widest scope, is about (i) 2D constraints across multiple views (matching constraints), (ii) characterizations of the space of all images of a particular object (indexing functions). In other words, one seeks to best represent, in terms of efficiency, compactness, flexibility and scope of use, two kinds of manifolds: (i) the manifold of image and viewing parameters (invariance to shape), and (ii) the manifold of image and object parameters (invariance to viewing parameters).

A further distinction of this line of research is that one generally prefers to find linear functions that describe the inherently non-linear relationship between 3D objects and their 2D views. One can argue whether for a particular application, say landscape reconstruction from aerial photographs, it really matters if the final computational method is linear or non-linear (especially when a good initial guess can be obtained for the numerical solution), but in general, seeking ways to embed the non-linearities into a space where the manifold in question becomes a linear subspace constitutes an important step forward in understanding and manipulating visual information.

On the photometric domain an issue of interest is how to combine brightness information from images, in the form of spatial and temporal derivatives, with geometric constraints arising from the fact that all images are coming from the same 3D static scene. This issue is relevant to the problem of recovering correspondences from a sequence of images, and to the problem of recovering scene structure and observer motion parameters *directly* from brightness measurements. An important step in achieving these goals is being able to recast the general motion estimation problem into a parametric framework, i.e., that each pixel contributes measurements to a fixed number of unknowns. The latter implies a process of elimination (either structure of motion parameters), which brings us back to the geometric goals of embedding the non-linear view-manifold in spaces where it becomes a linear subspace — as we shall see in the sequel.

Finally, the meeting point between geometry and photometry comes from considering lines in addition to points. A line can be interpreted as coming from a 3D line (a geometric entity), or as an uncertainty caused by insufficient brightness measurements (the so called "aperture" problem) — a photometric entity.

We will show in this paper that common to all these goals is the so called "trilinear tensor".

1.1 Background on the Trilinear Tensor

The first work to combine points and lines into one framework is due to Aloimonos & Spetsakis [18]. They have shown that three views admit a set 27 numbers arranged in three matrices. Matching points were shown to produce three linear equations for the 27 numbers (hence 9 points are required), and matching lines produce two equations (hence 13 lines are required). They have also addressed the connection between lines and uncertainty by showing that the aperture problem which is inherent in two views does not exist with three views. This line of research continued with Shashua [13] by attempting to generalize the "linear combination" result of Ullman & Basri [20], showed that the point geometry of three views actually produces 4 linearly independent equations, each having a trilinear form in image coordinates, for a tensor of 27 numbers (which happen to be the same numbers of [18]). These equations were called "trilinearities". Thus 7 matching points across three views are sufficient for linearly generating the intrinsic geometry of three views. In that work the equation relating the tensor and the (projective) motion parameters was first derived. Hartley [5] rederived the tensor equation but with a different indexing scheme (adopted later in this paper), and Shashua & Werman [17] have shown that certain rearrangements of the tensor elements (equivalent to certain contractions of the tensor) produce *collineations* of the 2D plane, i.e., projective transformations due to intrinsic planes in space — a result which is important for unfolding the reconstruction problem in a very simple manner and for the question of multiple (>3) views, as described in the sequel.

The trilinearities were rederived by Faugeras & Mourrain [4] using exterior algebra. That had two distinct advantages: first, a geometric interpretation was given to the trilinear equations, second, the method of derivation was simple and general and which led later to further work by Weinshall, Werman & Shashua [21] and Carlsson [2] on "dual" tensors. Similarly, the trilinearities were rederived by Triggs [19], using Penrose tensorial notations, and Heyden [7], which together with [4, 22], established the existence of quadlinear forms (with total of 81 coefficients) across four views with the negative result that further views would not add any new constraints.

2 The Trilinear Tensor

Consider two perspective views ψ, ψ' of a 3D scene. Let P be a point in 3D projective space projecting onto matching points $p \in \psi, p' \in \psi'$ in 2D projective plane. The relationship between the 3D and 2D spaces is represented by the 3×4 matrices, $[I, 0]$, $[A, v']$, i.e.,

$$\begin{aligned} p &= [I,0]P \\ p' &\cong [A, v']P \end{aligned} \qquad (1)$$

We may adopt the convention that $p = (x, y, 1)^\top$, $p' = (x', y', 1)^\top$, and therefore $P = (x, y, 1, \rho)$. The coordinates $(x, y), (x'y')$ are matching points (with respect to some arbitrary image origin — say the geometric center of each image plane). The vector v' is the translational component of camera motion and is the view of the center of projection of the first camera in view ψ'. The matrix A is a 2D projective transformations (collineation, homography matrix) from ψ to ψ' induced by *some* plane in space (the plane $\rho = 0$). In a calibrated camera setting the plane $\rho = 0$ is the plane at infinity and A is the rotational component of camera motion and $\rho = 1/z$ where z is the depth of the point P in the first camera coordinate frame. For more details on the representation and methods for projective reconstruction see [3, 6, 12, 16, 11, 1].

Let s_k^l be the matrix,

$$s = \begin{bmatrix} -1 & 0 & x' \\ 0 & -1 & y' \end{bmatrix}$$

It can be verified by inspection that eqn. 1 can be represented by the following two equations (standard method for removing the scale in that equation):

$$\rho s_k^l {v'}^k + p^i s_k^l a_i^k = 0, \qquad (2)$$

with the standard summation convention that an index that appears as a subscript and superscript is summed over (known as a contraction). Superscripts denote contravariant indices (representing points in the 2D plane, like v') and subscripts denote covariant indices (representing lines in the 2D plane, like the rows of A). Thus, a_i^k is the element of the k'th row and i'th column of A, and ${v'}^k$ is the k'th element of v'. Note that we have two equations because $l = 1, 2$ is a free index.

Similarly, the camera transformation between views ψ and ψ'' is

$$p' \cong [B, v'']P.$$

Likewise, let r_j^m be the matrix,

$$r = \begin{bmatrix} -1 & 0 & x'' \\ 0 & -1 & y'' \end{bmatrix}$$

And likewise,

$$\rho r_j^m {v''}^j + p^i r_j^m b_i^j = 0, \qquad (3)$$

Note that k and j are dummy indices (are summed over) in equations 2 and 3, respectively. We used different dummy indices because now we are about to eliminate ρ and combine the two equations together. Likewise, l, m are free indices, therefore in the combination they must be separate indices. We eliminate ρ and obtain a new equation:

$$(s_k^l {v'}^k)(p^i r_j^m b_i^j) - (r_j^m {v''}^j)(p^i s_k^l a_i^k) = 0,$$

and after grouping the common terms:

$$s_k^l r_j^m p^i ({v'}^k b_i^j - {v''}^j a_i^k) = 0,$$

and the term in parenthesis is the trilinear tensor:

$$\boxed{\alpha_i^{jk} = {v'}^k b_i^j - {v''}^j a_i^k. \qquad i, j, k = 1, 2, 3} \qquad (4)$$

And the tensorial equations (the trilinearities) are:

$$\boxed{s_k^l r_j^m p^i \alpha_i^{jk} = 0}, \qquad (5)$$

Hence, we have four trilinear equations (note that $l, m = 1, 2$). In more explicit form, these functions (referred to as "trilinearities") are:

$$\begin{aligned}
x'' \boldsymbol{\alpha}_i^{13} p^i - x'' x' \boldsymbol{\alpha}_i^{33} p^i + x' \boldsymbol{\alpha}_i^{31} p^i - \boldsymbol{\alpha}_i^{11} p^i &= 0, \\
y'' \boldsymbol{\alpha}_i^{13} p^i - y'' x' \boldsymbol{\alpha}_i^{33} p^i + x' \boldsymbol{\alpha}_i^{32} p^i - \boldsymbol{\alpha}_i^{12} p^i &= 0, \\
x'' \boldsymbol{\alpha}_i^{23} p^i - x'' y' \boldsymbol{\alpha}_i^{33} p^i + y' \boldsymbol{\alpha}_i^{31} p^i - \boldsymbol{\alpha}_i^{21} p^i &= 0, \\
y'' \boldsymbol{\alpha}_i^{23} p^i - y'' y' \boldsymbol{\alpha}_i^{33} p^i + y' \boldsymbol{\alpha}_i^{32} p^i - \boldsymbol{\alpha}_i^{22} p^i &= 0.
\end{aligned}$$

Since every corresponding triplet p, p', p'' contributes four linearly independent equations, then seven corresponding points across the three views uniquely determine (up to scale) the tensor α_i^{jk}. More details and applications can be found in [13].

It readily follows from the fact that the tensor vanishes with the contraction of two covariant vectors and one contravariant vector that a matching triplet of a point and two lines provides one linear equation for the tensor elements. Likewise, two contractions with covariant vectors leaves us with a covariant vector, thus three matching lines provide two linear equations for the tensor elements. Finally, because a point is defined by the intersection of two lines, three matching points provide four linear equations for the tensor elements (as explicitly specified above).

3 Contractions, Collineations, Fundamental Matrix

Consider the contraction $e_j\alpha_i^{jk}$ where $e = (1,0,0)^\top$. The result is a 3×3 matrix, denoted by E_1. Similarly, let $E_2 = \alpha_i^{2k}$, and $E_3 = \alpha_i^{3k}$. We obtain a remarkably and simple result:

The three matrices E_1, E_2, E_3 are three homography matrices $E_j : \psi \mapsto \psi'$ of three distinct and intrinsic planes.

In other words, the three matrices are *collineations* of the 2D projective plane mapping points in ψ to points in ψ' induced by three distinct planes, respectively. The orientation and locations of each plane are determined by the motion parameters B, v''. More details and proofs can be found in [17].

Since 2D collineations are the building block for projective reconstruction, this result is very important. For example, the "fundamental" matrix F between ψ and ψ' can be linearly determined from the tensor by:

$$E_j{}^\top F + F^\top E_j = 0$$

which yields 18 linear equations of rank 8 for F. Similarly, cross products between columns of two collineations provide epipolar lines which can be used to recover the epipole v'. We will also use this property of the tensor to concatenate together multiple (> 3) views, as described in the sequel.

4 Tensor Brightness Constraint

Consider all lines s' in the 2D plane that are coincident with p', i.e., $s'_k p'^k = 0$. Note that the rows of the matrix s_k^l are covariant vectors (represent lines in the 2D plane), thus s' is spanned by the rows of s_k^l, and in turn equation (2) still holds:

$$\rho s'_k v'^k + p^i s'_k a_i^k = 0, \qquad (6)$$

for all covariant vectors s' spanned by by the two covariant vectors $(-1, 0, x')$ and $(0, -1, y')$ (representing vertical and horizontal lines, respectively).

In particular, consider the linear combination with coefficients I_x, I_y which are the components of the gradient vector ∇I measured at point (x, y) in the first view ψ. Thus,

$$s' = \begin{pmatrix} -I_x \\ -I_y \\ x'I_x + y'I_y \end{pmatrix}$$

Our next step is to remove the contribution of x', y' from s' (this is the only place in equation (6) where correspondence is required) and for that purpose we will use the "constant brightness equation" due to [8]:

$$I_x u + I_y v + I'_t = 0,$$

where $u = x - x'$, $v = y - y'$ and I'_t is the discrete temporal derivative at (x, y), i.e., $I_2(x, y) - I_1(x, y)$ where I_2 and I_1 are the image intensity values of the second and first image, respectively.

After substituting the constant brightness equation in s' we obtain:

$$s' = \begin{pmatrix} -I_x \\ -I_y \\ I'_t + xI_x + yI_y \end{pmatrix} \qquad (7)$$

Thus equation (6) with the covariant vector s' given by (7) is a brightness constraint equation that relates camera motion A, v', object shape ρ and brightness information in the form of spatial and temporal derivatives. The situation so far is not very different from that of Horn & Weldon [9] with the difference that here we are using an uncalibrated (projective) model, rather than the Longuett-Higgins & Prazdny [10] model of small motion, and we use a different style of notations.

The next step is identical in concept to the way the trilinear tensor was derived above. We eliminate the contribution of object shape ρ by using the third view ψ''. The constant brightness equation between the first and third view becomes: $I_x(x - x'') + I_y(y - y'') + I''_t = 0$, and by going through exactly the same steps as before we obtain:

$$\rho s''_j v''^j + p^i s''_j b_i^j = 0, \qquad (8)$$

and

$$s'' = \begin{pmatrix} -I_x \\ -I_y \\ I''_t + xI_x + yI_y \end{pmatrix} \qquad (9)$$

We eliminate ρ and obtain a new equation, the "Tensor Brightness Constraint":

$$\boxed{s'_k s''_j p^i \alpha_i^{jk} = 0} \qquad (10)$$

The new constraint relates the observer motion and certain measurable quantities (products of elements of s' and s'') of image spatial and temporal derivatives. Because the structure of the world does not appear in the constraint equation we have a *parametric* constraint, which means that *every* pixel with non-vanishing spatial gradient contributes an equation with a *fixed* number of unknowns. Moreover, the constraint equation is *linear* in the elements of α_i^{jk}. Therefore, in principle, we can obtain a linear least-squares solution for the elements of α_i^{jk} just by measuring spatial and temporal derivatives. More details can be found in [15].

5 Rank 4: Properties of the Tensor Manifold

The ultimate goal is to find the means for combining together the contribution of any number of views, not only three. To this end we suggest to start with investigating the space of *all* trilinear tensors and look for rank deficiencies in that space. Any finding of that sort is extremely useful because it readily allows a statistical way of putting together many views, simply by means of factorization. The main result is that trilinear tensors across $m > 3$ views are embedded in a low dimensional linear subspace.

Consider the following arrangement: we are given views $\psi_1, \psi_2, ...\psi_{m+2}$, $m \geq 1$. For each (ordered) triplet of views there exists a unique trilinear tensor. Rather than considering all triplets of views, we consider the m triplets that contain ψ_1, ψ_2, i.e., the triplets $< \psi_1, \psi_2, \psi_i >$, $i = 3, ..., m+2$. Consider each of the tensors as a vector of 27 components (arrange the components arbitrarily, but stick with this arrangement for all tensors) and concatenate all these vectors as columns of a $27 \times m$ matrix. The question is what is the rank of this matrix when $m \geq 27$? Clearly, if the rank is smaller than 27 we obtain a line of attack on the task of putting together many views. The motivation for considering this arrangement is that a view adds only 12 parameters (up to scale). It may be the case that the redundancy of representing an additional view with 27 numbers (a column vector in the $27 \times m$ matrix), instead of 12, comes to bear only at a non-linear level — in which case it will not affect the rank of the system above. Therefore, a rank deficiency implies an important property of a collection of tensors. We can prove the following result:

All trilinear tensors live in a manifold of $\mathcal{P}^{26}$. The space of all trilinear tensors with two of the views fixed, is a 12'th dimensional linear sub-space of $\mathcal{R}^{27}$.

Therefore, the rank is 12, thus each additional view adds, linearly, only 12 parameters — as expected. An immediate consequence of this result is:

A linear combination of tensors $< \psi_1, \psi_2, \psi_i >$ and $< \psi_1, \psi_2, \psi_j >$ produces an admissible tensor $< \psi_1, \psi_2, \psi >$, for some view ψ.

The corollary is not as obvious as it may seem. All tensors live in a non-linear manifold because there are algebraic dependencies among the tensor elements. Thus, the line passing through two arbitrary points on that manifold does not necessarily live inside the manifold. The fact that it does for the selection of points described in the corollary, is therefore, not obvious. One application, for instance, of the corollary is view synthesis and animation.

A similar result applies to the space of *all* collineations (homography matrices) between two fixed views. Given some plane in space projecting onto views ψ and ψ', the corresponding image points are mapped to each other by a collineation (homography matrix), $Ap \cong p'$ for all matching pairs p, p'. Since the homography matrix A depends on the orientation and location of the planar object, we obtain a *family* of homography matrices when we consider all possible planes.

Consider homography matrices $A_1, A_2, ..., A_k$ each as a column vector in a $9 \times k$ matrix. We ask again, what is the rank of the system? It would be convenient if it were 4, because each additional homography matrix represents a plane, and a plane is determined by 4 parameters. We can prove the following result:

The space of all homography matrices between two fixed views is embedded in a 4 dimensional linear subspace of $\mathcal{R}^9$.

We can combine the rank 4 result with the result described previously that a tensor can be contracted into three homography matrices and obtain a "rank 4" result on the space of tensors, as follows. We recall from Section 3 that the tensor α_i^{jk} can be contracted into three homography matrices, associated with three distinct planes, between ψ and ψ'. Hence, consider the same situation as before where we have views $\psi_1, \psi_2, ...\psi_{m+2}$ and consider the tensors of the triplets $< \psi_1, \psi_2, \psi_i >$, $i = 3, ..., m+2$. But now, instead of arranging each tensor as a 27 column vector, we arrange it in a 9×3 block, where each column is the homography α_i^{jk}, $j = 1, 2, 3$. We obtain a $9 \times 3m$ matrix. Its rank must be 4:

A tensor of views $< \psi_1, \psi_2, \psi_3 >$ and "third" of the tensor $< \psi_1, \psi_2, \psi_4 >$, linearly span, with 12 coefficients, all tensors $< \psi_1, \psi_2, \psi >$ (over all views ψ). Each such tensor can be recovered using 6 matching points with ψ_1 and ψ_2.

We can improve and get even a tighter result on the minimal information required to linearly span the family of tensors $< \psi_1, \psi_2, \psi >$ (for all views ψ). Assume we have the fundamental matrix F and epipole v'. It is known from the work of [11] that the matrix $[v']_x F$, where $[v']_x$ is the skew-symmetric matrix associated with vector products, is a homography matrix (linearly independent from the three homography matrices provided by the tensor). We have therefore the following result:

The tensor of views $< \psi_1, \psi_2, \psi_3 >$ and the epipolar constraint (matrix F and epipole v') together lin-

early span, with 12 coefficients, all other tensors $<\psi_1, \psi_2, \psi>$ (running over all views ψ). Each additional view ψ contributes linearly 12 parameters and its tensor with ψ_1, ψ_2 can be determined linearly using 6 matching points.

More details on the material presented in this section can be found in [14].

6 summary

This paper has presented results on the goal of capturing the inter-relationship, geometrically and photometrically, across multiple perspective views. The main analysis vehicle is the "trilinear tensor" which captures in a very simple and straightforward manner the basic structures associated with this problem of research. We have not described in detail the particular applications to reconstruction, recognition and animation, but these can be found in some of the references provided in the text.

References

[1] P.A. Beardsley, A. Zisserman, and D.W. Murray. Navigation using affine structure from motion. In *Proceedings of the European Conference on Computer Vision*, pages 85–96, Stockholm, Sweden, May 1994.

[2] S. Carlsson. Duality of reconstruction and positioning from projective views. In *Proceedings of the workshop on Scene Representations*, Cambridge, MA., June 1995.

[3] O.D. Faugeras. Stratification of three-dimensional vision: projective, affine and metric representations. *Journal of the Optical Society of America*, 12(3):465–484, 1995.

[4] O.D. Faugeras and B. Mourrain. On the geometry and algebra of the point and line correspondences between N images. In *Proceedings of the International Conference on Computer Vision*, Cambridge, MA, June 1995.

[5] R. Hartley. Lines and points in three views — a unified approach. In *Proceedings Image Understanding Workshop*, Monterey, CA, November 1994.

[6] R. Hartley. Projective reconstruction and invariants from multiple images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(10):1036–1040, 1994.

[7] A. Heyden. Reconstruction from image sequences by means of relative depths. In *Proceedings of the International Conference on Computer Vision*, pages 1058–1063, Cambridge, MA, June 1995.

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

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

[10] 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.

[11] Q.T. Luong and T. Vieville. Canonic representations for the geometries of multiple projective views. In *Proceedings of the European Conference on Computer Vision*, pages 589–599, Stockholm, Sweden, May 1994. Springer Verlag, LNCS 800.

[12] A. Shashua. Projective structure from uncalibrated images: structure from motion and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):778–790, 1994.

[13] A. Shashua. Algebraic functions for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):779–789, 1995.

[14] A. Shashua and S. Avidan. The rank 4 constraint in multiple (≥ 3) view geometry. Technical report, Technion, CS Dept., October 1995.

[15] A. Shashua and K.J. Hanna. The tensor brightness constraints: Direct estimation of motion revisited. Technical report, Technion, CS Dept., October 1995.

[16] A. Shashua and N. Navab. Relative affine structure: Theory and application to 3D reconstruction from perspective views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 483–489, Seattle, Washington, 1994.

[17] A. Shashua and M. Werman. On the trilinear tensor of three perspective views and its underlying geometry. In *Proceedings of the International Conference on Computer Vision*, June 1995.

[18] M.E. Spetsakis and J. Aloimonos. A unified theory of structure from motion. In *Proceedings Image Understanding Workshop*, 1990.

[19] B. Triggs. Matching constraints and the joint image. In *Proceedings of the International Conference on Computer Vision*, pages 338–343, Cambridge, MA, June 1995.

[20] S. Ullman and R. Basri. Recognition by linear combination of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13:992—1006, 1991. Also in M.I.T AI Memo 1052, 1989.

[21] D. Weinshall, M. Werman, and A. Shashua. Shape tensors for efficient and learnable indexing. In *Proceedings of the workshop on Scene Representations*, Cambridge, MA., June 1995.

[22] M. Werman and A. Shashua. Elimination: An approach to the study of 3D-from-2D. In *Proceedings of the International Conference on Computer Vision*, June 1995.