GENERALIZED CAYLEY-CHOW COORDINATES AND COMPUTER VISION

BRIAN OSSERMAN

Abstract. A fundamental problem in computer vision is to reconstruct the configuration of a collection of cameras from the images they have taken of a common subject. If we model a camera as a linear projection from projective 3-space to the projective plane, this problem can be rephrased algebrogeometrically in terms of recovering a subvariety of a product of projective planes. Both the equations defining these subvarieties and the relevant Hilbert scheme were studied in work of Aholt, Sturmfels and Thomas in 2011. We explain how various mathematical tools can both give new explanations for known phenomena in computer vision, and lead to some new results. The most substantive mathematical contribution is a theory of Cayley-Chow coordinates for subvarieties of products of projective spaces, generalizing the usual case of a single projective space. We find that there are nontrivial dimensional inequalities on when this generalization works, and that these inequalities explain the existence and non-existence of the "multifocal tensors" from computer vision.

This is joint work (very much still in progress) with Matthew Trager.

1. Camera configurations and multiview varieties

One of the basic problems in computer vision involves reconstructing a collection of unknown camera positions from the resulting images. This is a central topic in the field, and one of the oldest, with very practical applications: for instance, it is now possible for a computer to use a collection of photos pulled from the internet to reconstruct a 3D model of famous landmarks, and even cities. However, recently this problem has been recast from an algebrogeometric perspective, leading to the potential for new advances coming from tools of algebraic geometry.

The basic classical model for a camera is as a linear projection from the three-dimensional world (considered as a $\mathbb{P}^3$) to the two-dimensional film/sensor plane (considered as a $\mathbb{P}^2$). Thus, an n-tuple of (positioned) cameras corresponds to an n-tuple of linear projections, which together induce a rational map

$$\mathbb{P}^3 \dashrightarrow (\mathbb{P}^2)^n.$$

The closure of the image of this map is thus a three-dimensional subvariety of $(\mathbb{P}^2)^n$, which is called the **multiview variety**. This can be thought of as describing which n-tuples of points in $\mathbb{P}^2$ could come from a single point in $\mathbb{P}^3$. Knowing the multiview variety is equivalent to knowing the camera configuration, at least up to change of 'world coordinates' in $\mathbb{P}^3$.

A typical approach to determining camera configurations is to look at data coming from two cameras (the 'fundamental matrix') or three cameras (the 'trifocal tensor'), which can be thought of as imposing bilinear or trilinear equations vanishing on the multiview variety. There is also a quadrifocal tensor encoding configurations of four cameras, but it has long been known that there are no tensors encoding configurations of n cameras, for $n > 4$. The

original motivation for my work with Trager on the theory of multigraded Cayley-Chow coordinates was to explain this phenomenon.

In a parallel direction, we have been thinking about how many equations are necessary in order to uniquely determine a multiview variety. Aholt, Sturmfels and Thomas computed the multidegree of a multiview variety, and one idea we have been exploring is to use intersection theory and the theory of multidegrees to show that under certain circumstances, even though a collection of $2n-3$ equations doesn't cut out the multiview variety precisely (and may even have excess-dimensional components), it still can only contain a unique multiview variety. On the other hand, motivated by thinking about the moduli space of multiview varieties (also studied by Aholt, Sturmfels, and Thomas) we have discovered that multiview varieties can in find be determined by fewer than $2n-3$ equations in general.

2. Multigraded Cayley-Chow coordinates

With the aforementioned motivation, we now discuss a purely mathematical construction (still joint with Trager). Let $X \subset \mathbb{P}^n$ be a projective variety of dimension r and degree d. The set of all linear spaces of dimension $n-r-1$ meeting X is a hypersurface Z_X in the Grassmannian $G(n-r-1,n)$. Any such hypersurface can be written as the zero set inside the Grassmannian of a polynomial F_X in the Plucker coordinates, which turns out to also be of degree d. The polynomial F_X is known as the **Chow form** or **Cayley form** of X, and we will refer to it as the Cayley-Chow form. From the Cayley-Chow form F_X we immediately recover the hypersurface Z_X, and one can then recover X as the set of points $P \in \mathbb{P}^n$ such that every $(n-r-1)$-dimensional linear space containing P corresponds to an element of Z_X.

Thus, the Cayley-Chow form can be used to encode subvarieties of projective space, and this classical construction has played an important role in moduli space theory, especially in the guise of Chow varieties, but also for instance in Grothendieck's original construction of Quot schemes.

Typically, if one wants to consider subvarieties of a different variety Y, one imbeds Y into projective space and then applies the above construction. But this loses any information about how the subvariety lies with respect to Y. We will consider how to generalize the theory of Cayley-Chow forms to subvarieties of a product of projective spaces without reimbedding, thereby retaining more information. To do this, we replace the linear spaces in the classical case with products of linear spaces, which opens up some flexibility. However, it turns out that for the theory to go through, certain constraints will have to be satisfied.

Recall that the Chow ring of $\mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_m}$ has a basis consisting of products of linear spaces, which means that it is isomorphic as a ring to $\mathbb{Z}[t_1,\ldots,t_m]/(t_1^{n_1+1},\ldots,t_m^{n_m+1})$, where each t_i is the class of the preimage of a hyperplane in the ith space. Thus, in this representation the Chow class of a subvariety of codimension c is homogeneous polynomial of degree c, which is also called the **multidegree**. Given $\gamma_1,\ldots,\gamma_m$ adding up to c, the coefficient of $t_1^{\gamma_1}\cdots t_m^{\gamma_m}$ in the multidegree can be determined by intersecting with the product of a general m-tuple of linear spaces, with the ith one having codimension γ_i.

We then have the following, which we are fairly confident is correct, but are in the process of working out the proof of.

Almost-Theorem 2.1. *(O.-Trager) Let $X \subseteq \mathbb{P}^{n_1} \times \cdots \times \mathbb{P}^{n_m}$ be a subvariety of dimension r, with multidegree $\sum c_{\gamma_\bullet} t_1^{\gamma_1} \cdots t_m^{\gamma_m}$. Fix $\beta_1, \ldots, \beta_m$ with $0 \leqslant \beta_i \leqslant n_i$ for each i, and*

$$\sum_i \beta_i = r + 1,$$

and suppose further that for each $I \subsetneq \{1, \ldots, m\}$, we have

$$\sum_{i \in I} \beta_i \leqslant \dim p_I(X), \tag{2.1.1}$$

where $p_I(X)$ is projection onto the coordinates in I. Let $Z_X \subseteq \prod_{i=1}^m \mathbb{G}(n_i - \alpha_i, n_i)$ consist of those $(L_1, \ldots, L_m)$ such that $X \cap (L_1 \times \cdots \times L_m) \neq \emptyset$, where $\mathbb{G}(n_i - \alpha_i, n_i)$ is the Grassmannian of $(n_i - \alpha_i)$-dimensional projective linear subspaces of $\mathbb{P}^{n_i}$.

Then Z_X is a hypersurface, and X can be recovered from Z_X. Moreover, Z_X can be realized as the zero set of a single multihomogeneous polynomial in the m sets of Plucker coordinates, and the multidegree of this polynomial is determined by the coefficients $c_{\gamma_\bullet}$ where $\gamma_\bullet$ varies over sequences obtained by adding 1 to exactly one term of $n_1 - \beta_1, \ldots, n_m - \beta_m$.

Conversely, if (2.1.1) is not satisfied, then either Z_X is not a hypersurface, or X cannot be recovered from Z_X.

Remark 2.2. The condition of (2.1.1) can only be satisfied if $m \leqslant r + 1$, since otherwise the set of i such that $\beta_i \neq 0$ is necessarily proper in $\{1, \ldots, m\}$, and will violate the necessary inequality. In the computer vision situation, this means that we must have $m \leqslant 4$, which explains the lack of multifocal tensors for more than four cameras.

Remark 2.3. According to recent work of Castillo, Li and Zhang, the support of the multidegree of X is determined by the collection of $\dim p_I(X)$. We have checked that conversely the $\dim p_I(X)$ can be determined by the support of the multidegree, and are in the process of working out how (2.1.1) can be expressed in terms of the multidegree.

Remark 2.4. Beyond the relevance to multifocal tensors, we hope that our construction will be of use in constructing new invariants of configurations of generalized algebraic cameras, as studied recently by Ponce-Sturmfels-Trager and Escobar-Knutson.

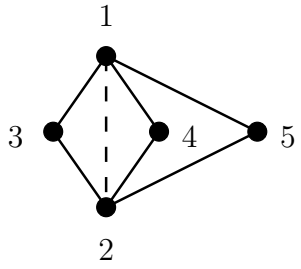
The proof of Almost-Theorem 2.1 (insofar as we've worked it out) involves incidence correspondence arguments to show that Z_X is a hypersurface and that X can be recovered from Z_X. This is more interesting than the classical case because Z_X does not recover X directly, but rather recovers X together with some extraneous components. However, we can recognize X among these components because we know its multidegree.

To show that Z_X is given by a polynomial we are led to consider the interesting question of when a tensor product of UFDs is still a UFD. This is not obvious in general (the unique factorization property is notoriously finicky), but we find that with some mild hypotheses on units and universality, Gauss' argument for polynomial rings goes through for much more general tensor products, giving us what we want.

3. Recovering camera configurations

In a different direction, we describe some of our work more directly on the algebraic vision side. If we have two cameras, the multiview variety has codimension 1 in $\mathbb{P}^2 \times \mathbb{P}^2$, so the fundamental matrix gives a single bilinear equation which cuts out the multiview variety. In the general case, we often attempt to repeatedly consider pairs of cameras at a time to determine the multiview variety. It is known that although the multiview variety is not cut out by $2n-3$ such bilinear equations, it can nonetheless be determined by an appropriate choice of $2n-3$ equations (provided that the camera centers are not all on a line, in which case it is known that one cannot recover the camera configuration even using all pairs of cameras). We reinterpret this in terms of multidegrees, analyzing the intersection of the $2n-3$ bilinear equations directly. We show that – even though the intersection may have components of dimension strictly greater than 3 – the intersection contains a unique component which can contain (and is in fact equal to) a multiview variety. Given that there are many problems in computer vision involving determination of how much data is necessary to recover a configuration of cameras, we hope that these techniques will be useful to recover new results as well. We plan to investigate this specifically in the aforementioned case of generalized algebraic cameras, although it could potentially apply to various more classical situations as well.

In a complementary direction, the moduli space of camera configurations has dimension $11n-15$, and the fundamental matrix from a pair of cameras imposes 7 conditions. Thus, it seems naively that one might be able to uniquely determine a configuration using fewer than $2n-3$ pairs of cameras. We can view this in terms of starting with a graph with no edges and n vertices, and inserting edges one at a time until we have enough information to determine the rest of the camera configuration. The third leg of a triangle can only impose 4 conditions, since the moduli space is 18-dimensional for $n=3$, so we want to avoid making triangles. The first example in which this leads to the possibility of fewer than $2n-3$ pairs is for $n=5$, where we have indeed found the following case:



Here we check that we can determine the configuration with the shown 6 edges, rather than the $7 = 2n-3$ previously described. Interestingly, this means that there is a unique multiview variety in the intersection of the 6 bilinear equations, even though every component of this intersection has dimension at least 4.

We can understand what is going on here geometrically. Let $P_1, \ldots, P_m$ be the camera centers in $\mathbb{P}^3$, and let $\mathbb{P}^2_i$ denote the $\mathbb{P}^2$ which is the image resulting from the ith camera. We can think of the fundamental matrix between the ith and jth cameras as giving us two pieces of information:

(1) points $P_{i,j} \in \mathbb{P}^2_i$ and $P_{j,i} \in \mathbb{P}^2_j$ which are the images of P_j and P_i respectively;
(2) an identification of the $\mathbb{P}^1$ of lines through $P_{i,j}$ with the $\mathbb{P}^1$ of lines through $P_{j,i}$.

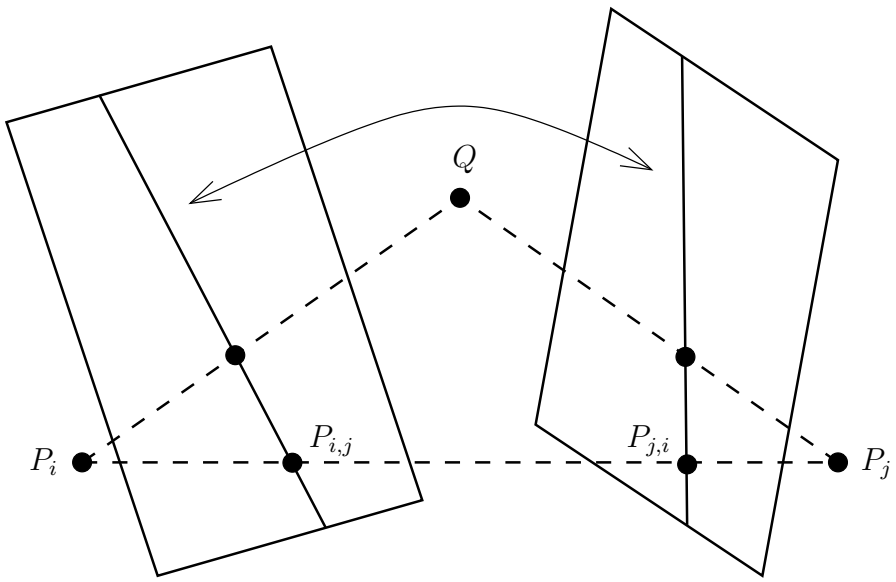


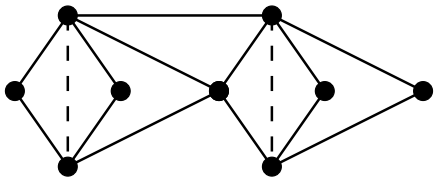
FIGURE 1. Identification of lines in image planes for a pair of cameras.

The latter arises as follows: if we have $Q \in \mathbb{P}^3$ not lying on the line through P_i and P_j, then Q, P_i, P_j span a plane, and this plane projects onto a line through $P_{i,j}$ in $\mathbb{P}^2_i$ and a line through $P_{j,i}$ in $\mathbb{P}^2_j$. These lines will be identified with one another. See Figure 1.

Now, suppose we have fundamental matrices for (i, j) and for (j, k); how far are we from having the fundamental matric for (i, k)? (Note that this will determine the camera triple (i, j, k)) The point $P_{j,i}$ is the image of P_i in $\mathbb{P}^2_j$; we don't have enough information to find $P_{k,i} \in \mathbb{P}^2_k$, but at least by considering the line through $P_{j,i}$ and $P_{j,k}$, we obtain a line $L_{k,i}$ in $\mathbb{P}^2_k$ which $P_{k,i}$ must lie on. Similarly, we obtain a line $L_{i,k}$ in $\mathbb{P}^2_i$ which $P_{i,k}$ must lie on. Finally, we note that the isomorphism of lines through $P_{i,k}$ with lines through $P_{k,i}$ must map $L_{i,k}$ to $L_{k,i}$, because both contain the image of P_j. This explains geometrically why the third fundamental matrix only gives 4 conditions: one each for finding $P_{i,k}$ on $L_{i,k}$ and $P_{k,i}$ on $L_{k,i}$, and two for giving an isomorphism of two $\mathbb{P}^1$s each of which already has one point marked.

Now, if we consider the 5-vertex figure above, under suitable generality hypotheses, for $i = 3, 4, 5$, considering the edges from 1 to i and from i to 2 will give us three lines on which $P_{1,2}$ must lie and three lines on which $P_{2,1}$ must lie, overdetermining those two points. They will also give us three marked points on the $\mathbb{P}^1$s of lines through these points, precisely determining the isomorphism. Thus, from this we can recover the $(1, 2)$ fundamental matrix, and it is standard that once we have "triangulated" the graph in this way, we can recover the entire situation.

This example only reduces the number of fundamental matrix computations by 1, but can be repeated by adding four new vertices at a time, repeating the construction with a single overlap with the old vertex set, and adding in one new edge to complete a triangle.



In this way, we'll add 7 new edges for every four new vertices we'll add, meaning that we need roughly $\frac{7}{4}n$ fundamental matrices to find the configuration of cameras. Given that we starting off using roughly $2n$, and the best one could possibly hope for would be on the order of $\frac{11}{7}n$, this constitutes significant progress.