Joint Feature Distributions for Image Correspondence

Bill Triggs
CNRS-INRIA, 655 avenue de l'Europe, 38330 Montbonnot, France.
Bill.Triggs@inrialpes.fr ◇ *http://www.inrialpes.fr/movi/people/Triggs*

Abstract

We introduce 'Joint Feature Distributions', a general statistical framework for feature based multi-image matching that explicitly models the joint probability distributions of corresponding features across several images. Conditioning on feature positions in some of the images gives well-localized distributions for their correspondents in the others, and hence tight likelihood regions for correspondence search. We apply the framework in the simplest case of Gaussian-like distributions over the direct sum (affine images) and tensor product (projective images) of the image coordinates. This produces probabilistic correspondence models that generalize the geometric multi-image matching constraints, roughly speaking by a form of model-averaging over them. These very simple methods predict accurate correspondence likelihood regions for any scene geometry including planar and near-planar scenes, without ill-conditioning or explicit model selection. Small amounts of distortion and non-rigidity are also tolerated. We develop the theory for any number of affine or projective images, explain its relationship to matching tensors, and give results for an initial implementation.

Keywords: Joint Feature Distributions, matching constraints, multi-image geometry, feature correspondence, statistical modelling.

1 Introduction

This paper introduces a natural statistical framework for multi-image feature correspondence, **Joint Feature Distributions (JFD's)**, and uses them to "probabilize" the entire range of affine and projective geometric matching constraints [12, 15, 2, 3, 5, 7, 21, 20]. JFD's are simply joint probability distributions over the positions of corresponding features in $m > 1$ different images, used as a summary of some population of interesting correspondences (all valid ones, those near a particular surface or object, background ones...). Conditioning on some of the features gives tight probabilistic correspondence search regions for the remaining ones. Although we will choose parametric forms that reproduce and generalize the standard matching constraints, JFD's are in essence descriptive statistical models rather than normative geometric ones: they aim to summarize the observed behaviour of the given training correspondences, not to rigidly constrain them to an ideal predefined geometry. We believe that JFD's will become the standard method for many correspondence search problems. Their benefits over matching constraints include: more precise search focusing; built-in handling of noise and distortion (small non-rigidities, lens distortion...); and globally stable estimation, even for geometries that are degenerate for classical matching constraints. The projective two image model is perhaps the most useful. It generalizes the epipolar constraint, but instead of searching along the full length of epipolar lines, it searches ellipses (Gaussians) whose centre, axis, length and width are determined by the point being matched, its epipolar line, the range of disparities seen in the training data, and the noise level. JFD's stably and accurately adapt to any scene geometry from deep 3D through to coplanar: as the depth range of the training data decreases, the search ellipses progressively shorten until ultimately the model becomes essentially homographic. There is no ill-conditioning for near-coplanar scenes, no need to choose between epipolar and homographic correspondence models, and no under- or over-estimation of the plausible correspondence regions. In contrast, epipolar models typically search entire (and perhaps inaccurate) epipolar lines, wasting effort and greatly increasing the probability of false correspondences.

The idea of using JFD's for image correspondence is very natural and obvious in retrospect. As far as we know it has not appeared before, but there are many related threads in the literature. JFD's react against explicit model selection for matching constraints [9, 10, 17] by incorporating the well-known statistical rule that you can only predict events similar to the ones you trained on — extrapolating a full epipolar geometry from near-coplanar data is unstable, but irrelevant for predicting near-coplanar correspondences, *c.f. e.g.* [13]. JFD's have analogies with Bayesian model averaging [18] but are much simpler and more direct. Plane+parallax [11, 14, 8, 23, 1, 22] offers stabler geometric parametrizations than matching tensors for near-planar scenes. These could no doubt be "probabilized" in much the same way as we do here.

Another aspect of this work is a new theoretical framework for studying multi-image geometry and especially matching constraints, based on the notion of the **tensor joint image**. We will use some isolated results from this,

Extended version of a paper appearing in the 2001 IEEE Int. Conf. Computer Vision. © 2001 IEEE Computer Society Press. This work was supported by European Union FET project VIBES.

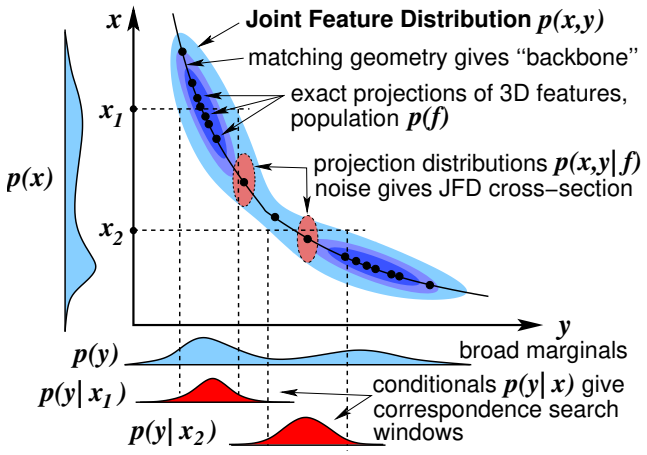

Figure 1: The basic principle of JFD feature correspondence.

but the full development had to be omitted for lack of space.

§2 sketches the general principles of JFD matching, §3 develops some tools, then we focus on Gaussian-like JFD models for affine (§4) and perspective (§5) camera geometries. §6 briefly discusses the implementation and some preliminary experiments and §7 concludes.

Notation: We assume familiarity with affine and projective matching constraints at the level of, *e.g.* [6], and the ability to think tensorially at need [21,20]. Slanted fonts denote inhomogeneous $\boldsymbol{x},\boldsymbol{y}$ and homogeneous $\boldsymbol{x},\boldsymbol{y}$ image vectors, upright fonts inhomogeneous $\mathbf{x},\mathbf{y}$ and homogeneous x,y 3D ones. P denotes 3×4 image projection matrices, $\mathbf{p}()$ probability distributions, $[\,\cdot\,]_\times$ 3×3 cross product matrices ($[\,\boldsymbol{x}\,]_\times\,\boldsymbol{y}=\boldsymbol{x}\wedge\boldsymbol{y}$), $\langle-\rangle_{\boldsymbol{x}}$ expectation over the distribution of $\boldsymbol{x}$.

2 Joint Feature Distributions

We can model the m noisy image projections $\boldsymbol{x}_{i\,|\,i=1\ldots m}$ of a fixed 3D feature $\mathbf{f}$ as probability distributions $\mathbf{p}_i(\boldsymbol{x}_i\,|\,\mathbf{f})$ centred on $\mathbf{f}$'s true projections, with widths determined by the relevant noise levels. More generally, the joint distribution $\mathbf{p}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\,|\,\mathbf{f})$ is typically well-localized, and for independent noise factors as $\prod_i \mathbf{p}_i(\boldsymbol{x}_i\,|\,\mathbf{f})$. If $\mathbf{f}$ now varies across some population of 3D features with distribution $\mathbf{p}(\mathbf{f})$, the **Joint Feature Distribution (JFD)** of the resulting population of image features is:

$$\mathbf{p}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m) \;\equiv\; \int \mathbf{p}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\,|\,\mathbf{f})\,\mathbf{p}(\mathbf{f})\,\mathbf{df} \qquad (1)$$

For broad priors $\mathbf{p}(\mathbf{f})$, the one-image marginals $\mathbf{p}(\boldsymbol{x}_i) \equiv \int \mathbf{p}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m)\,\boldsymbol{dx}_1\ldots\boldsymbol{dx}_{i-1}\boldsymbol{dx}_{i+1}\ldots\boldsymbol{dx}_m = \int \mathbf{p}_i(\boldsymbol{x}_i\,|\,\mathbf{f})\,\mathbf{p}(\mathbf{f})\,\mathbf{df}$ are typically broad and uninformative. But the 'sharpness' of the original image projections $\mathbf{p}_i(\boldsymbol{x}_i\,|\,\mathbf{f})$ is not entirely lost: *The JFD* $\mathbf{p}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m)$ *remains highly correlated and still encodes most of the precise location information.* In particular, the **Conditional Feature Distributions (CFD's)** like

$$\mathbf{p}(\boldsymbol{x}_1\,|\,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m) \;\equiv\; \frac{\mathbf{p}(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m)}{\mathbf{p}(\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m)} \qquad (2)$$

encode precise inter-image dependencies that are efficient tools for correspondence search. Fig.1 illustrates the principle for two 1D projective images $\boldsymbol{x},\boldsymbol{y}$ of a 1D scene. The JFD $\mathbf{p}(\boldsymbol{x},\boldsymbol{y})$ encodes a strong probabilistic dependency between $\boldsymbol{x}$ and $\boldsymbol{y}$ whose "backbone" is the underlying geometric correspondence (here a 1D homography). The marginals $\mathbf{p}(\boldsymbol{x})$, $\mathbf{p}(\boldsymbol{y})$ are broad and uninformative, but given a particular feature $\boldsymbol{x}$, the CFD $\mathbf{p}(\boldsymbol{y}\,|\,\boldsymbol{x})$ (the normalized cross-section through $\mathbf{p}(\boldsymbol{x},\boldsymbol{y})$ at $\boldsymbol{x}$) is sharply peaked at the corresponding $\boldsymbol{y}$ value. We can use this to predict tight probabilistic search regions for $\boldsymbol{y}$ given $\boldsymbol{x}$, and vice versa.

The abstract JFD framework has rich analogies with, and generalizes, conventional multi-image geometry — see table 1. It applies to any correspondence relationship that can be modelled probabilistically, regardless of feature type, parametrization, number of images, rigidity or distortion. But it is most useful when (suitable parametric forms can be chosen so that) the estimated JFD's have strong correlations that provide accurate search focusing. The link with geometry is strongest for correspondences governed by matching constraints. Then, as in fig.1, the matching geometry forms the "backbone" of the JFD $\mathbf{p}(\boldsymbol{x},\boldsymbol{y})$ and fixes the locations of the CFD's $\mathbf{p}(\boldsymbol{y}\,|\,\boldsymbol{x})$, while the image noise determines the cross-section of the JFD and the widths of the CFD's. The 3D feature population $\mathbf{p}(\mathbf{f})$ or its image marginals $\mathbf{p}(\boldsymbol{x}),\mathbf{p}(\boldsymbol{y})$ determine the height of the JFD along its backbone, but have little direct influence on the CFD shapes.

As with matching constraints, JFD's are image-based models originally derived from 3D quantities (here the 3D feature prior $\mathbf{p}(\mathbf{f})$ and the projection models $\mathbf{p}_i(\boldsymbol{x}_i\,|\,\mathbf{f})$, there the camera matrices P_i), but typically estimated from observed image correspondences. The familiar three stage estimation process [4] still applies: (*i*) build a large set of possible correspondences, *e.g.* by feature detection followed by correlation matching; (*ii*) hypothesize well-supported candidate models, *e.g.* using a robust clusterer such as RANSAC; (*iii*) robustly fit parametric model(s) to the most interesting candidate(s). The fitted models are parametric probability distributions for both JFD's (explicitly) and matching constraints (we actually fit a geometry-based probabilistic noise model). The clustering stage ensures reliable fitting by rejecting false matches and 'uninteresting' true ones, *e.g.* features on moving objects when we are fitting the background, or non-coplanar features when we are fitting a plane. Good clustering is even more critical for JFD's than for matching constraints owing to their polymorphism: they are designed to summarize a user-defined class of observations not to enforce

Entity	**Matching Constraint Approach**	**Joint Distribution Approach**
3D camera geometry	Camera projection mapping, matrices P_i: $\mathbf{f} \rightarrow x_i = \mathsf{P}_i \mathbf{f}$	Conditional feature projection distributions $\mathbf{p}(\boldsymbol{x}_i \,|\, \mathbf{f})$
Image signature of camera geometry	Multi-image matching tensors $T_{ij\ldots k}$	Joint Feature Distributions $\mathbf{p}(\boldsymbol{x}, ..., \boldsymbol{z})$
Inter-image feature transfer	Tensor based feature transfer $x \simeq T_{ij\ldots k} \cdot y \cdot \ldots \cdot z$	Conditional Feature Distributions $\mathbf{p}(\boldsymbol{x} \,|\, \boldsymbol{y}, ..., \boldsymbol{z})$
Inter-image feature correspondence	Geometric matching constraints $T_{ij\ldots k} \cdot x \cdot \ldots \cdot z = 0$	Probability that features correspond, $\mathbf{p}(\boldsymbol{x}, ..., \boldsymbol{z})$, or $\mathbf{p}(\boldsymbol{x} \,|\, \boldsymbol{y}, ..., \boldsymbol{z})$
Scene reconstruction	Ray intersection, tensor-based reconstruction	Posterior 3D feature probability $\mathbf{p}(\mathbf{f} \,|\, \boldsymbol{x}, ..., \boldsymbol{z})$

Table 1: Analogies between the joint distribution approach and multi-image matching constraints.

a predefined structure, so it is much less clear what constitutes an outlier. The obvious approach is to use self-consistency, finding clusters too dense to be probable under broader members of the parametric distribution family *c.f. e.g.* [4, 18]. We will not go into these difficult grouping issues here, but we expect JFD's to be effective correspondence models for many natural grouping classes such as points on compact moving objects.

From now on we focus on deriving efficient parametric models for JFD's. We will only consider Gaussian-like models, which appear to be the simplest useful parametric forms. Gaussians can only capture linear dependencies, so to produce JFD's that can mimic the standard matching constraints, we will need parametrizations that make these constraints appear linear. As in matching constraint estimation, we do this by mapping the input observations into a suitable *joint image* space, containing the direct sum (juxtaposition) of the input coordinates for affine models, but their tensor (outer) product for projective ones. For example, for projective fundamental matrix or epipolar JFD estimation, we map the n correspondences $(\boldsymbol{x}, \boldsymbol{x}')$ to homogeneous 9-D outer product vectors $x \otimes x' \sim (1, x, y, x', y', xx', xy', yx', yy')^\top$ and build a $9 \times n$ measurement matrix M from these. The fundamental matrix estimate uses just the smallest eigenvector of $\mathsf{M}\mathsf{M}^\top$ (*e.g.* [6]), whereas the JFD model captures the underlying uncertainty using an appropriately-weighted "average" over all of the eigenvectors (in fact, $(\mathsf{M}\mathsf{M}^\top)^{-1}$). Conditioning the JFD gives compact correspondence search regions consistent with all (not just one!) of the likely models in the average. The JFD is loosely analogous to "model averaging" of fundamental matrices [18], but it is based directly on the input correspondences, not on a blurred geometric model.

3 Scatter & Covariance

Homogeneous covariance: Before starting we introduce some tools. We encode distributions homogeneously. Given an uncertain affine point $\boldsymbol{x}$ with mean $\bar{\boldsymbol{x}}$, covariance $\boldsymbol{V}$ and homogeneous vector $x = \binom{\boldsymbol{x}}{1}$, its **homogeneous covariance** X, **homogeneous information** X^{-1} and χ^2 **value** are:

$$X \equiv \langle x\, x^\top \rangle_{\boldsymbol{x}} = \begin{pmatrix} \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\top + \boldsymbol{V} & \bar{\boldsymbol{x}} \\ \bar{\boldsymbol{x}}^\top & 1 \end{pmatrix} \quad (3)$$

$$X^{-1} = \begin{pmatrix} \boldsymbol{V}^{-1} & -\boldsymbol{V}^{-1}\bar{\boldsymbol{x}} \\ -\bar{\boldsymbol{x}}^\top \boldsymbol{V}^{-1} & +\bar{\boldsymbol{x}}^\top \boldsymbol{V}^{-1}\bar{\boldsymbol{x}} + 1 \end{pmatrix} \quad (4)$$

$$\chi^2(\boldsymbol{x} \,|\, \bar{\boldsymbol{x}}, \boldsymbol{V}) \equiv (\boldsymbol{x} - \bar{\boldsymbol{x}})^\top \boldsymbol{V}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}) = x^\top X^{-1} x - 1 \quad (5)$$

The mean, covariance and information of the Gaussian fit neatly into the homogeneous matrices X, X^{-1}. Given a collection of training points $\{\boldsymbol{x}_p\}_{p=1\ldots n}$, their **homogeneous scatter matrix** $\frac{1}{n}\sum_p x_p\, x_p^\top$ encodes their mean and covariance, and hence defines an approximate Gaussian probability model for the point population. If the points are also uncertain, their **smoothed homogeneous scatter** $\frac{1}{n}\sum_p \langle x_p\, x_p \rangle_{\boldsymbol{x}_p}$ encodes the mean and covariance of the mixture distribution generated by the sum of the individual point distributions. Viewed as a summary of the population statistics, this double-counts the noise and hence overestimates the covariance, but when there are relatively few points this smoothed but biased estimate is often preferable to the unsmoothed one because it contains additional information about the noise level. Either type of scatter matrix can be used when estimating JFD's below.

Note that these formulae require the homogeneous point vectors to be affinely normalized (scaled so that the last coordinate is 1). We assume this throughout the paper. Although many formulae (notably in §5) appear projective, they all are based on the standard "noise in pixels" image plane error model which is intrinsically affine. For a start at building a projectively covariant error model, see [19].

Dual covariance: Some matching constraints are based on the lines through an image point rather than the point itself. The simplest is the homographic relationship $x \simeq H y$. This can be written in constraint form as $u H y = 0 = v H y$ where u, v are any two independent lines through x. For least squares estimation we square and sum the constraints: $0 \approx (u H y)^2 +$

$(v\,H\,y)^2 = y^\top H^\top (u\,u^\top + v\,v^\top)\,H\,y$. We will view $u\,u^\top + v\,v^\top$ as the "homogeneous scatter matrix" of the chosen set $\{u, v\}$ of lines through x. More generally we could use $\frac{2}{n}\sum_{i=1}^n u_i\,u_i^\top$ where $\{u_i\}$ is any rank 2 set of lines through x. These rank 2 matrices encode x as their null vector, and each defines its own importance weighting over the lines through x (*i.e.* the constraints). To be more systematic, we fix a standard weighting procedure that defines our notion of "the uniform distribution of lines" through any given x. Algebraically, the most uniform way to write the squared homography constraints is $\|\,[\,x\,]_\times\,H\,y\|^2 \approx 0$, which leads to the "scatter matrix" $[\,x\,]_\times^\top\,[\,x\,]_\times = \|x\|^2 I - x\,x^\top$. This is not projectively covariant, but we can make it so by introducing a fixed quadric matrix Q, which we usually take to be the identity matrix in a well-normalized projective frame. We then define the **dual covariance** of x to be $\widetilde{X} \equiv (x^\top Q\,x)\,Q - (Q\,x)(Q\,x)^\top$. Or if x is uncertain with homogeneous covariance $X = \langle x\,x^\top\rangle$, the dual covariance is the expectation of this: $\widetilde{X} \equiv \mathrm{trace}(Q\,X)\,Q - Q\,X\,Q$. For $Q = I$ this becomes (*c.f.* (3)):

$$\widetilde{X} = \begin{pmatrix} 1{+}y^2{+}V_{yy} & -xy{-}V_{xy} & -x \\ -xy{-}V_{xy} & 1{+}x^2{+}V_{xx} & -y \\ -x & -y & x^2{+}y^2{+}V_{xx}{+}V_{yy} \end{pmatrix} \quad (6)$$

For nonsingular Q, $X = \frac{1}{d}\,\mathrm{trace}(Q^{-1}\,\widetilde{X})\,Q^{-1} - Q^{-1}\,\widetilde{X}\,Q^{-1}$, so dualization is reversible (d is the space dimension, here 2). Our JFD models of line-through-point constraints (homographies, trifocal, quadrifocal) are all based on dual covariances.

4 The Affine JFD

We now develop a Gaussian JFD model for point features under affine image projection. This is the simplest useful model, and a good warm-up for the projective case. To keep the projective link clear we work in affine-homogeneous coordinates rather than centred inhomogeneous ones. Our JFD model must reproduce and "probabilize" the affine matching constraints. But these are already linear in the image coordinates and Gaussians naturally model linear relationships, so the problem is trivial. Suppose that the training data is n correspondences in m affine images $\boldsymbol{x}_{ip\,|\,i=1\ldots m,p=1\ldots n}$. Collect the components of each correspondence into a $2m+1$ component homogeneous **affine joint image** vector $\mathsf{x}_p \equiv (\boldsymbol{x}_{1p}^\top \ldots \boldsymbol{x}_{mp}^\top\ 1)^\top$, and form these into a $(2m+1)\times n$ **affine measurement matrix** $\mathsf{M} \equiv (\mathsf{x}_1, \ldots, \mathsf{x}_n)$. Viewing our correspondences as points in joint image space, their homogeneous scatter matrix is simply $\mathsf{V} \equiv \frac{1}{n}\sum_{p=1}^n \mathsf{x}_p\,\mathsf{x}_p^\top = \frac{1}{n}\,\mathsf{M}\,\mathsf{M}^\top$, or if we choose to use the smoothed scatter: $\mathsf{V} \equiv \frac{1}{n}\sum_p \langle \mathsf{x}_p\,\mathsf{x}_p^\top\rangle = \frac{1}{n}\sum_p \left(\mathsf{x}_p\,\mathsf{x}_p^\top + \left(\begin{smallmatrix} \boldsymbol{V}_p & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{smallmatrix}\right)\right) = \frac{1}{n}\,\mathsf{M}\,\mathsf{M}^\top + \frac{1}{n}\left(\begin{smallmatrix} \sum_p \boldsymbol{V}_p & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{smallmatrix}\right)$, where $\boldsymbol{V}_p$ is the $2m\times 2m$ inhomogeneous joint noise covariance of x_p (for independent noise, $\boldsymbol{V}_p$ is block diagonal with 2×2 blocks). Our affine JFD model is simply the Gaussian that best describes this population of joint image vectors, *i.e.* the Gaussian with homogeneous covariance V and homogeneous information V^{-1}.

Why does this work? - The theory of affine projection tells us that for ideal noiseless observations, $\mathrm{rank}(\mathsf{M}) \leq 4$, *i.e.* the vectors x_p span a 3D affine space [16]. We are modelling noisy observations, but the ideal behaviour tells us what to expect: M typically has 1 large 'homogenization', ≤ 3 large 'geometry' and $\geq 2m-4$ small 'noise' singular vectors, and similarly for $\mathsf{V} = \frac{1}{n}\mathsf{M}\,\mathsf{M}^\top$ with eigenvectors. So the JFD is typically very 'flat' — broad and featureless along the 'geometry' directions, but narrow along the remaining 'noise' ones. Conditioning on an image point $\boldsymbol{x}$ effectively freezes two of the 'geometry' directions, so at most one remains, spanning the joint epipolar line of $\boldsymbol{x}$ in the remaining images. Even this direction is restricted to the breadth of the training population, so for coplanar data it will shrink to a point.

For our Gaussian JFD's, conditioning leads to familiar Schur-complement matrix formulae. To do the calculation, partition $(\boldsymbol{x}_1^\top \ldots \boldsymbol{x}_m^\top)^\top$ into known components $\boldsymbol{k}$ and unknown ones $\boldsymbol{u}$, freeze $\boldsymbol{k}$ at their known values in $\chi^2(\mathsf{x}) = \mathsf{x}^\top\mathsf{V}^{-1}\mathsf{x} - 1$ (5), and complete the squares to find the conditional log likelihood of the remaining unknowns $\boldsymbol{u}$. Let $\bar{\boldsymbol{k}}, \bar{\boldsymbol{u}}$ be the training set means of $\boldsymbol{k}, \boldsymbol{u}$. Partition the corresponding information as $\left(\begin{smallmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{smallmatrix}\right)$, where $\boldsymbol{A} = (\mathsf{V}^{-1})_{\boldsymbol{kk}}$, *etc*. The search region for the unknowns $\boldsymbol{u}$ given the knowns $\boldsymbol{k}$ is defined by the CFD $\mathrm{p}(\boldsymbol{u}\,|\,\boldsymbol{k})$, which turns out to have mean $\bar{\boldsymbol{u}} - \boldsymbol{C}^{-1}\boldsymbol{B}^\top(\boldsymbol{k} - \bar{\boldsymbol{k}})$ and covariance $\boldsymbol{C}^{-1}$. (NB: The *population* covariance of $\boldsymbol{u}$ is $\mathsf{V}_{\boldsymbol{uu}} = (\boldsymbol{C} - \boldsymbol{B}^\top\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}$, which is usually much larger). If $\boldsymbol{k}$ is an uncertain measurement with covariance $\boldsymbol{D}^{-1}$, the CFD $\mathrm{p}(\boldsymbol{u}\,|\,\boldsymbol{k})$ is broadened to mean $\bar{\boldsymbol{u}} - \boldsymbol{C}^{-1}\boldsymbol{B}^\top(\boldsymbol{A}+\boldsymbol{D} - \boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^\top)^{-1}\boldsymbol{D}(\boldsymbol{k} - \bar{\boldsymbol{k}})$ and covariance $(\boldsymbol{C} - \boldsymbol{B}^\top(\boldsymbol{A}+\boldsymbol{D})^{-1}\boldsymbol{B})^{-1}$. (Usually $\boldsymbol{D} \gg \boldsymbol{A}$ so we can drop the $\boldsymbol{A}$'s). The prior likelihood for observing the knowns $\boldsymbol{k}$ in the first place is $\chi^2(\boldsymbol{k}\,|\,\mathsf{V}) = (\boldsymbol{k} - \bar{\boldsymbol{k}})^\top(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^\top)(\boldsymbol{k} - \bar{\boldsymbol{k}})$. As usual in such calculations, there are other forms for these expressions that may be stabler or more efficient, but we will not go into this here.

Implementation is straightforward: form the homogeneous scatter V from the training data, invert to get the information V^{-1} (the Gaussian JFD model), partition and condition on known observations to get search windows for their unknown correspondents. One minor snag is that $\mathsf{V} = \mathsf{M}\,\mathsf{M}^\top$ becomes rank deficient (rank ≤ 4) for noiseless data, so the estimated information V^{-1} becomes infinite. This is correct — exact geometry allows infinitely accurate predictions — but numerically inconvenient. In practice we avoid it by adding a small diagonal regularizer $\mathrm{diag}(\epsilon, \ldots, \epsilon, 0)$ to V before inverting. Typically $\epsilon \sim 10^{-9}$:

large enough to prevent loss of numerical precision during the inversion, but not so large as to blur the final estimates significantly. Similarly, V is rank deficient for $n \leq 2m$ noisy but unsmoothed correspondences because we do not have enough observations to estimate all of the noise covariances. The solution is to incorporate more noise information, *e.g.* using smoothed scatters.

5 The Projective JFD

Affine JFD's are too rigid to model perspective distortion exactly, so we now develop more flexible projective models. As before we consider only Gaussian-like models, so to mimic the matching constraints we need to use parametrizations in which these become linear. Projective matching constraints are *multi*linear in the homogeneous coordinates of their image features, but as in "linear" matching tensor estimation we can make the problem appear linear by treating multilinear combinations as if they were independent coordinates, *i.e.* by mapping the input feature vectors to their outer (Kronecker) product tensor. For example, for two images we can not use just the image coordinates $\mathsf{x} = (x, y, 1)^\top, \mathsf{x}' = (x', y', 1)^\top$ or the affine joint image vector $(x, y, x', y', 1)^\top$, because the projective matching constraints also have bilinear terms $xx', ..., yy'$. Instead we need to collect the components of the outer product $\mathsf{x}\,\mathsf{x}'^\top$ into a vector $(xx', xy', x, yx', yy', y, x', y', 1)^\top$ and use these as working coordinates. More generally, for point features in any number of images, it turns out (proof omitted) that tensoring the input coordinates is necessary and sufficient to linearize all of the matching constraints linking the images, and generically the only linear constraints on these coordinates are matching ones. So we really have no choice: to linearize the projective matching constraints linking features $\mathsf{x}_1, ..., \mathsf{x}_m$ from m images, we have to use the 3^m components of their **joint image tensor** $\mathsf{t} = \mathsf{x}_1 \otimes ... \otimes \mathsf{x}_m$ as working coordinates. We will view t both as a 3^m-component vector and as a tensor $\mathsf{t}^{A...D} = \mathsf{x}_1^A \cdot ... \cdot \mathsf{x}_m^D$ (indices $A...D = 1...3$). Assuming affine normalization for $\mathsf{x}_1, ..., \mathsf{x}_m$, our projective JFD models are "Gaussians in t-space", $\mathbf{p}(\mathsf{t}) \sim e^{-L/2}$, with negative log likelihood:

$$L = \mathsf{t}^\top \mathsf{W}\, \mathsf{t} = \mathsf{W}_{A...D\,A'...D'} (\mathsf{x}_1^A ... \mathsf{x}_m^D)(\mathsf{x}_1^{A'} ... \mathsf{x}_m^{D'}) \tag{7}$$

The JFD is parametrized by the **homogeneous information tensor** W, viewed as a symmetric positive definite $3^m \times 3^m$ matrix generalizing the homogeneous information. This model has the following useful properties:

1. It naturally models uncertain matching constraints. Algebraically, the simplest way to represent and combine uncertain constraints is to use weighted sums of squared constraint violations. For linear constraints such as the matching constraints on t, this yields nonnegative quadratic forms in the variables, *i.e.* Gaussians in t.
2. If we freeze some of the variables x_i at arbitrary values, L retains its tensored-quadratic form in the remaining ones, with coefficients given by W contracted against the frozen variables. So conditioning on known values for search region (CFD) prediction reduces to trivial tensor contraction — even simpler than the affine case.
3. Conditioning down to a single image gives a standard Gaussian expressed in homogeneous form, so predicting its high-probability search regions is easy.

JFD estimation: Now consider how to estimate a projective JFD W that summarizes a given set of training correspondences $(\mathsf{x}_{1p}, ..., \mathsf{x}_{mp})_{\,|\,p=1...n}$. By analogy with the affine case we treat the joint image tensors $\mathsf{t}_p = \mathsf{x}_{1p} \otimes ... \otimes \mathsf{x}_{mp}$ of the training correspondences as 3^m-component affine-homogeneous vectors and build their $3^m \times 3^m$ homogeneous scatter matrix $\mathsf{V} = \frac{1}{n}\sum_p \mathsf{t}_p\, \mathsf{t}_p^\top = \frac{1}{n}\mathsf{M}\,\mathsf{M}^\top$ where $\mathsf{M} = (\mathsf{t}_1, ..., \mathsf{t}_n)$ is the $3^m \times n$ measurement matrix familiar from linear matching tensor estimation. We then invert to get the JFD parameter estimate $\mathsf{W} \approx \mathsf{V}^{-1}$. This last step is unfortunately only heuristic[1] but it appears to work reasonably well in practice, perhaps because (if we imagine the eigen-decomposition of V being inverted) it gets at least the noise model in t's block of affine coordinates and the noiseless perspective corrections right.

As in the affine case, if we have uncertainties for the features we can use them to stabilize the JFD's noise level estimates, and we do this by taking expectations over noise when calculating the scatter. We assume independent noise so that tensor expectations factor into single-image ones. Working tensorially, the **smoothed scatter tensor** is:

$$\begin{aligned} \mathsf{V}^{A...D\,A'...D'} &= \tfrac{1}{n}\textstyle\sum_p \left\langle (\mathsf{x}_{1p}^A ... \mathsf{x}_{mp}^D)(\mathsf{x}_{1p}^{A'} ... \mathsf{x}_{mp}^{D'}) \right\rangle \\ &= \tfrac{1}{n}\textstyle\sum_p \left\langle \mathsf{x}_{1p}^A \mathsf{x}_{1p}^{A'} \right\rangle \cdot ... \cdot \left\langle \mathsf{x}_{mp}^D \mathsf{x}_{mp}^{D'} \right\rangle \\ &= \tfrac{1}{n}\textstyle\sum_p \mathsf{X}_{1p}^{AA'} \cdot ... \cdot \mathsf{X}_{mp}^{DD'} \end{aligned} \tag{8}$$

where $\mathsf{X}_{ip} = \left\langle \mathsf{x}_{ip}\, \mathsf{x}_{ip}^\top \right\rangle$ are the homogeneous covariances of the input features. Once again, this smoothes the JFD estimate at the cost of some double-counting of noise. It

[1] Our projective JFD's are not really Gaussians because their input coordinates t are restricted to a nonlinear $(2m)$-D subvariety of their (3^m)-D space — tensors of the rank-one form $\mathsf{t} = \mathsf{x}_1 \otimes ... \otimes \mathsf{x}_m$. Gaussian integrals over this restricted space are intractable, so we can not calculate the normalization factor that makes the JFD into a correctly normalized probability distribution. This factor is indispensable for estimating W. For training data with scatter V on a normalized distribution family $\mathbf{p}(\mathsf{t}) = e^{-(\mathsf{t}^\top \mathsf{W}\, \mathsf{t} - N(\mathsf{W}))/2}$, where $N(\mathsf{W})$ is the normalization, maximum likelihood estimation reduces to minimizing $\mathrm{trace}(\mathsf{W}\,\mathsf{V}) - N(\mathsf{W})$ with respect to W, with implicit solution W such that $\mathsf{V} = \frac{dN(\mathsf{W})}{d\mathsf{W}}$. For a true Gaussian, $N(\mathsf{W}) = \log\det \mathsf{W} - d\log(2\pi)$ and hence $\frac{dN}{d\mathsf{W}} = \mathsf{W}^{-1}$.

is particularly useful when there are $n < 3^m$ training features (which is common for $m \geq 3$). As a safeguard, we also add a $3^m \times 3^m$ diagonal regularizer $\mathrm{diag}(\epsilon, ..., \epsilon, 0)$ to V, where typically $\epsilon \sim 10^{-8}$. These measures are even more necessary in the projective case than in the affine one, as V is both large (so that many measurements are required to span it) and structurally ill-conditioned (because "perspective effects are usually small" compared to affine ones). The ill-conditioning is normal and causes no problems so long as we regularize enough to prevent it from causing loss of numerical precision.

'Epipolar' JFD *vs.* linear fundamental matrix estimation: Both methods start with the $9 \times n$ measurement matrix M of the tensored measurements. Form the 9×9 scatter $V = \mathsf{M}\mathsf{M}^\top$, let $V = \sum_{a=1}^{9} \lambda_a f_a f_a^\top$ be its eigendecomposition, and write the eigenvectors f_a as 3×3 "fundamental matrix candidates" F_a. The conventional linear fundamental matrix estimate is the smallest eigenvector F_9 of V, or equivalently the largest of $W = V^{-1}$. On test correspondences $t = x \otimes x'$, the JFD estimate $W = V^{-1}$ has unnormalized "log likelihood" penalty function $t^\top W t = \sum_{a=1}^{9} \lambda_a^{-1} (f_a\, t)^2 = \sum_{a=1}^{9} \lambda_a^{-1} |x\, F_a\, x'|^2$, *i.e.* "a weighted sum of possible epipolar constraints". Similarly, conditioning on x gives conditional log likelihood $x'^\top A\, x'$ for x', where $A_{A'B'} = W_{AB\,A'B'}\, x^A\, x^B$, *i.e.* the correspondence search regions are defined by "weighted scatters of possible epipolar lines" $A = \sum_{a=1}^{9} \lambda_a^{-1} (x^\top F_a)(x^\top F_a)^\top$. The fundamental matrix estimate amounts to truncating W at its largest eigenvector, giving effective penalty function $(f_9^\top t)^2 = |x\, F_9\, x'|^2$, *i.e.* the estimated epipolar constraint violation. For small noise and strong data, V has just one very small eigenvalue, so the penalty sum is entirely dominated by F_9 and the JFD model reduces to the fundamental matrix one. But if V has several small eigenvalues owing to noisy or weak data (*e.g.* coplanar data makes V rank 6 with 3 "noise" eigenvalues), these all contribute significantly the penalty sum, which becomes a kind of weighted average over these observed "constraints" on the data, restricting the directions in which the measurements can vary and hence the size of the "averaged epipolar" correspondence search regions.

Statistical error weighting: Although they include covariances, our linear JFD methods are essentially 'algebraic': they implement heuristic polynomial error models rather than statistically weighted rational ones. We will not consider nonlinear JFD estimation here, but a step towards statistical weighting in the conditioning calculation greatly improves the accuracy of the predicted search regions. For points near the epipole, algebraic weighting produces over-broad search regions (fig.2 top right). As above, the cost transversal to epipolar lines is controlled by the JFD's epipolar line violation term $|x\, F_9\, x'|^2$, where F_9 is the fundamental matrix. For x near the epipole, $x^\top F_9 \approx 0$ and $|x\, F_9\, x'|^2$ is small for any x'. We correct for this heuristically by replacing $A_{A'B'} = W_{AB\,A'B'}\, x^A\, x^B$ in the CFD $x'^\top A\, x'$ with λA where $\lambda = (W_{AB\,A'B'}\, V^{AB}\, N^{A'B'}) / \mathrm{trace}(A N)$, $N = \mathrm{diag}(1, 1, 0)$, and V is the x-image population scatter. The idea is that if $W \approx f_9 f_9^\top$ as above, $\mathrm{trace}(A N) \approx x^\top F_9 N F_9^\top x$ is the norm of the epipolar line vector $x^\top F_9$ and the numerator is the average of such norms across the training population. So λ reinforces the cost near the epipole without changing its overall population average. This heuristic reweighting procedure works well in practice and we are currently trying to formalize it.

5.1 Dual Space JFD's

The above approach in principle allows us to produce projective JFD's for any type of matching geometry in any number of images, and it is indeed the preferred representation for the practically important 2 image 'epipolar JFD' case. However, for $m > 2$ images or known-coplanar scenes it uses training correspondences very inefficiently. In the space of joint image tensors t, the nonlinear d-D image of a d-D projective subspace of 3D space containing k centres of projection turns out to span a $(\binom{m+d}{d} - k)$-D linear subspace (proof omitted). The above JFD model "learns" (spans) at most one tensor dimension per training correspondence, so just to capture the underlying geometry (let alone the noise) we need $\binom{m+3}{3} - m = 8, 17, 31...$ training correspondences in $m = 2, 3, 4...$ images for general 3D geometry, or $\binom{m+2}{2} = 6, 10, 15...$ for known-coplanar points. In comparison, linear matching tensor estimators need only $8, 7, 6$ correspondences for 3D points and $4, 4, 4$ for coplanar ones. Matching constraints are more efficient when they are tensor-valued, so that a single matching tensor with relatively few coefficients generates several linear constraints on each tensored image correspondence. By mirroring these index structures we can build correspondingly efficient JFD models, at the cost of less image-symmetric representations and an implicit restriction to correspondence models subjacent to the mirrored matching constraint. For example, a JFD based on the index structure of a two image homography constraint can be estimated from 4 correspondences rather than 6, but implicitly commits us to quasi-planar data.

To do this, we simply need to assume a JFD whose form is an average over constraints of the desired type. Free indices arise essentially when points x appear dualized as $[x]_\times$ in the matching constraints, *i.e.* for "any line through the point" style constraints. For example, 2 image homographic constraints with matrix H and 3 image trifocal constraints with tensor T can be written symbolically as $\| [x']_\times H x \|^2$ and $\| [x']_\times (T \cdot x) [x'']_\times \|^2$, or alternatively as $\sum_i |u_i'^\top H x|^2$ and $\sum_{ij} |(u_i')^\top (T \cdot x) u_j''|^2$ where $\{u_i'\}, \{u_j''\}$ are any sets of two (or

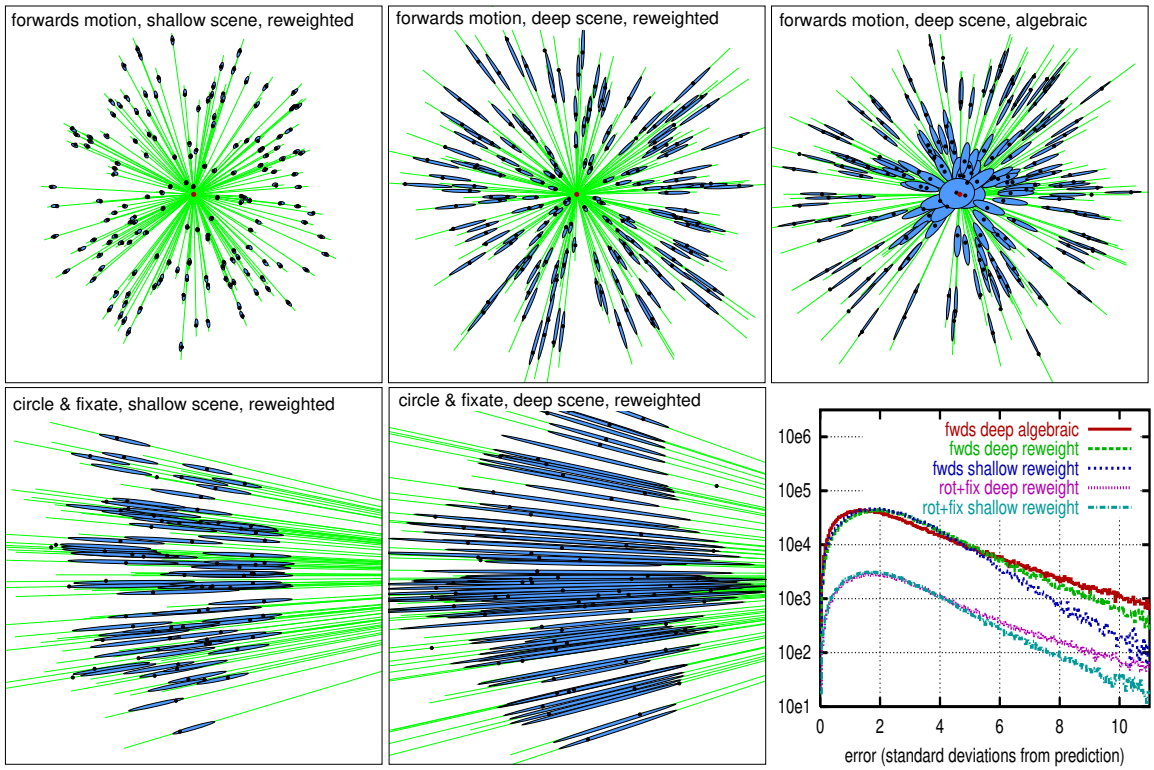

Figure 2: Predicted search ellipses for the projective 2 image JFD, versus the true correspondences and their epipolar lines. The scenes contain uniform random points in a sphere (deep scene) or a sphere flattened to 20% of its depth (shallow scene). *Top right:* Algebraic error weighting leads to incorrect broadening of the search regions near the epipole. The other plots use our CFD reweighting heuristic to correct this (see end of §5.0). *Bottom right:* For all of the geometries tested, the estimated conditional distributions accurately reflect the true correspondence likelihoods. Here we histogram the standard errors of the true observations under the estimated CFD's, for forwards (top curves) and fixation (bottom curves) motions.

more) independent lines through x', x''. Expanding these forms tensorially gives $(H^B_A\, H^{B'}_{A'}) X^{AA'}\, \widetilde{X}'_{BB'}$ and $(T^{BC}_A\, T^{B'C'}_{A'}) X^{AA'}\, \widetilde{X}'_{BB'}\, \widetilde{X}''_{CC'}$ where $X = x\,x^\top$ and $\widetilde{X}' = [x']^\top_\times [x']_\times$ or $\widetilde{X}' = \sum_i u'_i\, u'_i$, and similarly for $\widetilde{X}''$ with x'', u''_j. The squared constraints can be expressed compactly in terms of the homogeneous covariance X of x and the scatter matrices $\widetilde{X}', \widetilde{X}''$ of the lines through x', x''. We will fix the relative weighting of the different constraints by systematically using the dual covariances (6) of x', x'' for $\widetilde{X}', \widetilde{X}''$. For our JFD models we take averages over constraints of these forms, *i.e.* we adopt homogeneous Gaussian-like forms with unnormalized log likelihoods $W^{B\,B'}_{A\,A'}\, X^{AA'}\, \widetilde{X}'_{BB'}$ and $W^{BC\,B'C'}_{A\,A'}\, X^{AA'}\, \widetilde{X}'_{BB'}\, \widetilde{X}''_{CC'}$, where X and $\widetilde{X}', \widetilde{X}''$ are the (noiseless) normal and dual homogeneous covariances of the test correspondences (x, x', x''). These are still quadratic in the tensored measurements $x \otimes x' (\otimes x'')$, so they are reparametrizations of the general projective JFD models developed above. They are parametrized by information tensors $W^{B\,B'}_{A\,A'}$, $W^{BC\,B'C'}_{A\,A'}$, which can be viewed respectively as 9×9 and 27×27 homogeneous information matrices representing scatters $\sum_i h_i\, h_i^\top$ and $\sum_i s_i\, s_i^\top$ over possible homography matrices (9-component 'vectors' h_i) and trifocal tensors (27-component 'vectors' s_i). To estimate the models we again build (regularized and possibly smoothed) scatter tensors over training correspondences, here $V^{A\,A'}_{B\,B'} = \frac{1}{n}\sum_p X^{AA'}_p\, \widetilde{X}'_{p\,BB'}$ and $V^{A\,A'}_{BC\,B'C'} = \frac{1}{n}\sum_p X^{AA'}_p\, \widetilde{X}'_{p\,BB'}\, \widetilde{X}''_{p\,CC'}$, treat these as 9×9 and 27×27 homogeneous covariance matrices, and invert to estimate the corresponding information. To use the models for correspondence search, we condition on known feature positions by contracting their normal or dual covariances (as appropriate) into the information tensors until we reach a single-image model, which is necessarily quadratic in the remaining feature vector, *i.e.* a Gaussian whose likelihood regions are simple ellipses. This approach extends to all other matching constraint equations with free covariant indices: we just need to use normal or dual covariances as appropriate, and think tensorially when necessary.

6 Implementation & Experiments

We have implemented the above methods in MATLAB for any number of images and duality structure, but here we only show brief results for the 2 image 'epipolar' JFD. Recall the algorithm: build the 9×9 scatter matrix $V = \frac{1}{n}\sum_p t_p\, t_p^\top$ from the tensored training correspondences t_p, regularize and invert to get the JFD information $W = (V + \mathrm{diag}(\epsilon, ..., \epsilon, 0))^{-1}$, and view this as a tensor $W_{AB\,A'B'}$. Then, for each test correspondence x, form $A_{A'B'} = W_{AB\,A'B'}\, x^A x^B$ and optionally rescale it by $\lambda = (W_{AB\,A'B'}\, V^{AB} N^{A'B'})/\mathrm{trace}(AN)$ ($N = \mathrm{diag}(1, 1, 0)$) to correct the error weighting. The resulting A is the conditional information for $x' = (x', y', 1)^\top$, from which Gaussian log-likelihood search ellipses can be found by expanding $x'^\top A x'$ as a quadratic and discarding the constant term. Algebraic error weighting does produce over-wide search ellipses for points near the epipole, so it is advisable to include the reweighting factor λ. The reweighted method works well in practice for all of the geometries that we have tested, giving search ellipses aligned with the epipolar lines with realistic lengths and breadths, which progressively shrink to circles as the scene becomes planar.

7 Summary and Conclusions

We introduced Joint Feature Distributions (JFD's), a general statistical framework for image matching based on modelling the joint probability distributions of the positions of corresponding features in different images. The JFD is estimated from a population of training correspondences, then conditioned on the values of test features to produce tight likelihood regions for the corresponding features in other images. We developed relatively simple Gaussian-like JFD models for affine and projective images, which can be viewed as probabilistic "model averages" of the affine and projective multi-image matching constraints. The methods naturally and stably handle any scene geometry from deep 3D through to coplanar scenes, without explicit model selection. For example, the 'epipolar' JFD stably enforces an epipolar, homographic or near-homographic constraint, according to the behaviour of the training data.

Future work: The JFD idea is recent and we are still actively investigating its properties. There are some theoretical loose ends, particularly in the projective case where even the basic $W \approx V^{-1}$ estimation procedure for the "linear" model is only heuristic. We do not yet have JFD's with rigorous statistical error weighting, and it is unclear whether there are JFD analogues of matching tensor consistency relations like $\det(F) = 0$. Both issues are likely to lead to nonlinear models. Practically, we need to develop robust estimators for JFD's. As JFD's are less rigid than matching constraints, self-consistency based clustering will probably be needed to isolate correspondence subpopulations susceptible to JFD modelling. The full population model will thus be a mixture of JFD's. Numerically, we are developing QR and SVD based JFD representations that should be less sensitive to rounding errors than our current scatter / information ones.

A The Tensor Joint Image

This appendix develops the properties of tensor joint image representation. "Recall" two fundamental mappings from algebraic geometry. Given projective spaces $\mathbb{P}^A, ..., \mathbb{P}^D$ with generic points $x^A, ..., z^D$ and dimensions $d_1, ..., d_m$, the **Segré mapping** takes $(x^A, ..., z^D)$ in the Cartesian product (direct sum) space $\mathbb{P}^A \times ... \times \mathbb{P}^D$ to the rank one tensor[2] $t^{A...D} \equiv x^A \cdot ... \cdot z^D$ in the tensor product space $\mathbb{P}^{A...D}$. The **Segré variety** is the image of $\mathbb{P}^A \times ... \times \mathbb{P}^D$ under this mapping. It is a $(\sum_i d_i)$-dimensional algebraic variety in the $(\prod_i (d_i + 1) - 1)$-dimensional projective space $\mathbb{P}^{A...D}$, isomorphic to $\mathbb{P}^A \times ... \times \mathbb{P}^D$, and cut out by the 2×2 determinants of the form $t^{...A_i...A_j...}\, t^{...B_i...B_j...} - t^{...A_i...B_j...}\, t^{...B_i...A_j...} = 0$. Its points linearly span the whole of $\mathbb{P}^{A...D}$. The Segré mapping is the standard way of giving a Cartesian product a variety structure in algebraic geometry.

The Segré mapping *encapsulates the nonlinearity of multilinear polynomials on* $\mathbb{P}^A \times ... \times \mathbb{P}^D$, in the sense that any multilinear form $\sum_{A...D} c_{A...D}\, x^A ... z^D$ becomes a linear one $\sum_{A...D} c_{A...D}\, t^{A...D}$ in terms of the Segré coordinates $t^{A...D} = x^A ... z^D$. Any subvariety of $\mathbb{P}^A \times ... \times \mathbb{P}^D$ defined by multilinear polynomials is Segré-mapped to the intersection of a *linear* subspace (the null space of the Segré-linearized polynomials) with the Segré variety[3]. The individual homogeneous scale factors of $x^A, ..., z^D$ are confounded in $x^A \cdot ... \cdot z^D$, so the Segré mapping also turns out to be a good way of circumventing problems with homogeneous scale factors.

Similarly, given a projective space $\mathbb{P}^A$ with generic point x^A and dimension d, the **degree** m **Veronese mapping** on $\mathbb{P}^A$ takes x^A to the rank one tensor $x^{A_1 A_2 ... A_M} \equiv x^{A_1}\, x^{A_2} ... x^{A_m}$ in the symmetric tensor product space $\mathbb{P}^{(A_1 A_2 ... A_m)}$. The parentheses $(A_1 ... A_m)$ mean "take the symmetric part": in the tensor product it suffices to restrict attention to the $\binom{d+m}{d}$ *ordered* index combinations

[2] Several competing definitions of rank exist for tensors with more than 2 indices. None is entirely satisfactory, but all agree that outer products of vectors have rank 1, as here.

[3] Multilinear forms p in subsets of the variables $x^A, ..., z^D$ can be homogenized up to full multilinear forms by multiplying in turn by each multilinear combination $x^{A_i} \cdot ... \cdot y^{B_j}$ of the missing variables $x^{A_i}, ..., y^{B_j}$. Projectively, all entries of x^{A_i}, *etc.*, can not vanish at once, so the up-homogenized polynomials all vanish iff p does. In this projectivized sense, the Sergé mapping also linearizes multilinear polynomials of degree less than that of $x^A ... z^D$. The multi-image matching constraints behave this way.

$A_1 \leq A_2 \leq ... \leq A_m$ rather than the d^m unordered ones, as $\mathsf{x}^{A_1}\mathsf{x}^{A_2}...\mathsf{x}^{A_m}$ is automatically symmetric under arbitrary permutations of $A_1...A_m$. Analogously to the Segré case, the Veronese variety linearly spans $\mathbb{P}^{(A_1A_2...A_m)}$ and is cut out by 2×2 determinants, and the Veronese mapping *linearizes all degree m polynomials on* $\mathbb{P}^A$, mapping varieties defined by such polynomials (or, by uphomogenization, lower degree ones) to linear slices of the Veronese variety in $\mathbb{P}^{(A_1A_2...A_m)}$.

Now turn to vision. Consider m 3×4 image projections $\mathsf{P}^A_a, ..., \mathsf{P}^D_a$ projecting 3D points $\mathsf{X}^a \in \mathbb{P}^a$ to image points $\mathsf{x}^A \simeq \mathsf{P}^A_a \mathsf{X}^a, ..., \mathsf{z}^D \simeq \mathsf{P}^D_a$ in $\mathbb{P}^A, ..., \mathbb{P}^D$. Assemble the image points into a joint image[4] $(\mathsf{x}^A, ..., \mathsf{z}^D) \in \mathbb{P}^A \times ... \times \mathbb{P}^D$ — the image of $\mathbb{P}^a$ under the joint projection $(\mathsf{P}^A_a, ..., \mathsf{P}^D_a)$. A point-tuple is the image of some 3D point iff it satisfies certain well-known **geometric matching constraints** [12,15,2,3,5,7,21,20]. These constraints are multilinear in $(\mathsf{x}^A, ..., \mathsf{z}^D)$, so they become linear under the Segré mapping $(\mathsf{x}^A, ..., \mathsf{z}^D) \longrightarrow \mathsf{x}^A \cdot ... \cdot \mathsf{z}^D$, and hence cut out a linear-intersection subvariety of the Segré variety in the tensor product space $\mathbb{P}^{A...D}$. We call this the **tensor joint image**. It has coordinates of the form $t^{A...D} = (\mathsf{P}^A_a \mathsf{X}^a) \cdot ... \cdot (\mathsf{P}^D_d \mathsf{X}^d)$ and represents the image of $\mathbb{P}^a$ under the composition of joint image projection and Segré.

The image projections also act naturally on tensor products of $\mathbb{P}^a$, in particular taking a point (symmetric tensor) $\mathsf{T}^{a...d} \in \mathbb{P}^{(a...d)}$ to $t^{A...D} \simeq \mathsf{P}^A_a ... \mathsf{P}^D_d \, \mathsf{T}^{a...d}$ in the image tensor space $\mathsf{P}^{A...D}$. The image of the degree m Veronese mapping $\mathsf{X}^a \in \mathbb{P}^a \longrightarrow \mathsf{X}^a \cdot ... \cdot \mathsf{X}^d \in \mathbb{P}^{(a...d)}$ under this tensor product map is exactly the Segré mapping, so the Veronese variety of $\mathbb{P}^a$ also maps to the tensor joint image. In short, the following diagram is commutative:

$$\begin{array}{ccc} \mathbb{P}^a & \xrightarrow[\mathsf{X}^a \to \mathsf{X}^a \cdot ... \cdot \mathsf{X}^d]{\text{Veronese mapping}} & \mathbb{P}^{(a...d)} \\ \mathsf{X}^a \downarrow & & \mathsf{T}^{a...d} \downarrow \\ (\mathsf{P}^A_a \mathsf{X}^a, ..., \mathsf{P}^D_a \mathsf{X}^a) \quad \text{joint projection} & & \text{tensor projection} \quad \mathsf{P}^A_a ... \mathsf{P}^D_d \, \mathsf{T}^{a...d} \\ \mathbb{P}^A \times ... \times \mathbb{P}^D & \xrightarrow[(\mathsf{x}^A, ..., \mathsf{z}^D) \to \mathsf{x}^A \cdot ... \cdot \mathsf{z}^D]{\text{Segré mapping}} & \mathbb{P}^{A...D} \end{array}$$

The tensor projection $\mathsf{P}^A_a \cdot ... \cdot \mathsf{P}^D_d$ from the $\binom{m+3}{3}$-linear-dimensional space $\mathbb{P}^{(a...d)}$ to the (often much larger) 3^m-linear-dimensional one $\mathbb{P}^{A...D}$ is linear. Its kernel is spanned exactly by the m camera centre tensors $\mathbf{c}^a_i \cdot ... \cdot \mathbf{c}^d_i$, where $\mathbf{c}_i$ is the centre of projection of camera i. Generically the camera centre tensors are linearly independent in $\mathbb{P}^{(a...d)}$, so the image of $\mathbb{P}^{(a...d)}$ in $\mathbb{P}^{A...D}$ under the tensor projection generically has linear dimension $\binom{m+3}{3} - m$. The composite Veronese / tensor projection mapping has a base point at each camera centre, as expected.

Similarly, the Veronese images of 3D lines and planes have linear dimensions $\binom{m+1}{1}$ and $\binom{m+2}{2}$ in $\mathbb{P}^{(a...d)}$, and $\binom{m+1}{1} - k_1$, $\binom{m+2}{2} - k_2$ under tensor projection into $\mathbb{P}^{A...D}$, where k_1, k_2 are the number of camera centres they contain.

The Veronese image of $\mathbb{P}^a$ linearly spans $\mathbb{P}^{(a...d)}$, so its projection the tensor joint image spans the linear image of $\mathbb{P}^{(a...d)}$ in $\mathbb{P}^{A...D}$. The (Segré-mapped) matching constraints are simply the orthogonal complement of this linear subspace of $\mathbb{P}^{A...D}$. They can't be more restrictive without eliminating (the joint images of) real 3D points, and if they were less restrictive they would necessarily fail to eliminate some invalid image correspondences, as the Segré image of $\mathbb{P}^A \times ... \times \mathbb{P}^D$ linearly spans the whole of $\mathbb{P}^{A...D}$.

The point of all this is that tensoring the image measurements reduces much of the geometry of matching constraints to linear considerations (modulo the nonlinearity of the Segré mapping itself, of course). We claim that this is a good way to understand certain aspects of the structure of the matching constraints. In particular, it provides a suitable space in which to run a projective joint distribution formalism, as it allows simple Gaussian-like distributions to enforce the matching constraints.

[4] If we forget the individual projective depths (homogeneous scale factors) here we get the Cartesian-product joint image, if not we get the "projective joint image" defined in [21,20]. The latter is a *linear* image of $\mathbb{P}^a$ under the $3m \times 4$ "joint projection" matrix $(\mathsf{P}^A_a \, ... \, \mathsf{P}^D_a)^\top$, but not immediately recoverable from the input images. Either will do for our purposes as the Segré mapping below obliterates the relative scales.

References

[1] A. Criminisi, I. Reid, and A. Zisserman. Duality, rigidity and planar parallax. In *European Conf. Computer Vision*, pages 846–861. Springer-Verlag, 1998.

[2] O. Faugeras and B. Mourrain. On the geometry and algebra of the point and line correspondences between n images. In *Int. Conf. Computer Vision*, pages 951–6, 1995.

[3] O. Faugeras and T. Papadopoulo. Grassmann-Cayley algebra for modeling systems of cameras and the algebraic equations of the manifold of trifocal tensors. *Transactions of the Royal Society A*, 1998.

[4] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Computer Graphics and Image Processing*, 24(6):381–395, 1981.

[5] R. Hartley. Lines and points in three views and the trifocal tensor. *Int. J. Computer Vision*, 22(2):125–140, 1997.

[6] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[7] A. Heyden and K. Åström. A canonical framework for sequences of images. In *IEEE Workshop on Representations of Visual Scenes*, Cambridge, MA, June 1995.

[8] M. Irani and P. Anandan. A unified approach to moving object detection in 2D and 3D scenes. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 20(6):577–589, June 1998.

[9] K. Kanatani. *Statistical Optimization for Geometric Computation: Theory and Practice*. Elsevier Science, Amsterdam, 1996.

[10] K. Kanatani. Statistical optimization and geometric inference in computer vision. *Transactions of the Royal Society A*, 356(1740):1303–1320, 1998.

[11] R. Kumar, P. Anandan, and K. Hanna. Direct recovery of shape from multiple views: a parallax based approach. In *Int. Conf. Pattern Recognition*, pages 685–688, 1994.

[12] A. Shashua. Algebraic functions for recognition. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 17(8):779–89, 1995.

[13] A. Shashua and S. Avidan. On the reprojection of 3D and 2D scenes without explicit model selection. In *European Conf. Computer Vision*, pages 936–950, Dublin, 2000.

[14] A. Shashua and N. Navab. Relative affine structure: Canonical model for 3d from 2d geometry and applications. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 18(9):873–883, 1996.

[15] A. Shashua and M. Werman. On the trilinear tensor of three perspective views and its underlying geometry. In *Int. Conf. Computer Vision*, Boston, MA, June 1995.

[16] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Int. J. Computer Vision*, 9(2):137–54, 1992.

[17] P.H.S. Torr. Geometric motion segmentation and model selection. *Transactions of the Royal Society A*, 356(1740):1321–1340, 1998.

[18] P.H.S. Torr and A. Zisserman. Concerning bayesian motion segmentation, model averaging, matching and the trifocal tensor. In *European Conf. Computer Vision*, pages 511–527, Freiburg, 1998.

[19] B. Triggs. A fully projective error model for visual reconstruction. Unpublished. (Submitted to *ICCV'95 Workshop on Representations of Visual Scenes*).

[20] B. Triggs. The geometry of projective reconstruction I: Matching constraints and the joint image. Unpublished. (Submitted to *IJCV* in 1995).

[21] B. Triggs. Matching constraints and the joint image. In E. Grimson, editor, *Int. Conf. Computer Vision*, pages 338–43, Cambridge, MA, June 1995.

[22] B. Triggs. Plane + parallax, tensors and factorization. In *European Conf. Computer Vision*, pages 522–538, Dublin, 2000.

[23] D. Weinshall, P. Anandan, and M. Irani. From ordinal to euclidean reconstruction with partial scene calibration. In R. Koch and L. Van Gool, editors, *3D Structure from Multiple Images of Large-scale Environments SMILE'98*, pages 208–223. Springer-Verlag, 1998.