Multi-Frame Optical Flow Estimation Using Subspace Constraints

Michal Irani
Dept. of Computer Science and Applied Math
The Weizmann Institute of Science
76100 Rehovot, Israel

Abstract

We show that the set of all flow-fields in a sequence of frames imaging a rigid scene resides in a low-dimensional linear subspace. Based on this observation, we develop a method for simultaneous estimation of optical-flow across multiple frames, which uses these subspace constraints. The multi-frame subspace constraints are strong constraints, and replace commonly used *heuristic* constraints, such as spatial or temporal smoothness. The subspace constraints are geometrically meaningful, and are *not* violated at depth discontinuities, or when the camera-motion changes abruptly. Furthermore, we show that the subspace constraints on flow-fields apply for a *variety* of imaging models, scene models, and motion models. Hence, the presented approach for constrained multi-frame flow estimation is general. However, our approach does *not* require prior knowledge of the underlying world or camera model.

Although linear subspace constraints have been used successfully in the past for recovering 3D information (e.g., [18]), it has been assumed that 2D correspondences are given. However, correspondence estimation is a fundamental problem in motion analysis. In this paper, we use multi-frame subspace constraints to *constrain* the 2D correspondence estimation process itself, and *not* for 3D recovery.

1 Introduction

This paper presents an approach for simultaneous estimation of optical-flow across *multiple* frames. Optical flow (or "correspondence") estimation is usually applied to local image patches. Small regions, however, carry very little information (this is known as the "aperture problem"), and the optical flow estimates obtained are hence noisy and/or partial. To overcome this problem, spatial smoothness constraints are employed (e.g., [10, 1, 15]). However, these smoothness constraints are heuristic, and are violated especially at depth discontinuities. For a review and comparison of several of these optical flow techniques see [2]. Temporal smoothness constraints have also been introduced [5]. These, however, are violated when the camera motion changes abruptly.

Other methods overcome the aperture problem by applying *global* model constraints [7, 8, 3, 11, 17, 6, 4]. This allows the use of large analysis windows (often the entire image), which do not suffer from lack of local information. These techniques, however, assume an a-priori restricted model of the world or of the camera motion. For example, [11, 6, 4] assume a planar (or very distant) world. [7, 8, 17] assume a 3D world with dense 3D parallax, and will fail when applied to distant or planar worlds (which form a singular case for these algorithms). [3] reviews a hierarchy of such global motion models. While these methods perform well when the restricted model assumptions are applicable, they fail when these are violated.

Also, most methods for correspondence/flow estimation have been restricted to *pairs* of frames (or three frames [17]). With the rare exception of [8], most methods that use information from *multiple frames* rely on temporal smoothness. The resulting estimates are hence noisy and are "over-smoothed". In contrast, [8] exploits geometric consistency across multiple frames, but relies on prior knowledge that the underlying model is a 3D world with dense 3D parallax.

In this paper we develop an approach for simultaneously estimating correspondences across *multiple* frames by using information from all the frames, without assuming prior model selection. Our approach is based on the observation that the set of all flow-fields across multiple frames (that image the same scene) reside in a *low-dimensional linear subspace*. This is true despite the fact that different frames in the image sequence are obtained with different camera motions. The subspace constraints provide the additional constraints needed to resolve the ambiguity in image regions that suffer from the aperture problem. This is done *without* resorting to spatial or temporal smoothness. As opposed to smoothness constraints, the subspace constraints are geometrically meaningful, and are not violated at depth discontinuities or when camera-motion changes abruptly.

Linear subspace constraints have been used successfully in the past for recovering 3D information from *known* 2D correspondences (e.g., [18, 9]). In contrast,

we use multi-frame linear subspace constraints to *constrain* the 2D correspondence estimation process itself, without recovering any 3D information. Furthermore, we show that for a variety of world models (e.g., planar world vs. general 3D world) and a variety of camera models (e.g, orthographic vs. perspective cameras undergoing instantaneous motion) give rise to subspaces of very similar low dimensionalities. Because we employ subspace constraints based on the subspace *dimensionality* alone, these constraints can be used without prior knowledge of the underlying world or camera model.

In Sect. 2 we show that the set of all flow-fields across multiple frames (that image the same rigid scene) reside in a low-dimensional linear subspace. This is shown for a variety of motion models, scene models, and imaging models. In Sect. 3 we extend the multi-frame subspace constraints to apply directly to image brightness quantities. These are then incorporated in Sect. 4 into a simultaneous *multi-point multi-frame* flow algorithm, which takes advantage of the low-dimensionality subspace constraints within the estimation process itself. We conclude with some experimental results showing the benefits of the multi-frame constrained estimation.

2 Subspace Constraints on Flow-Fields

Let $I_1, ..., I_{\mathcal{F}}$ denote a sequence of $\mathcal{F}$ frames taken by a moving camera with *arbitrary* 3D motions. All frames are of the same size, and contain $\mathcal{N}$ pixels. Let I denote the *reference frame* in the sequence, i.e., the frame with respect to which all flow-fields will be estimated (e.g., the middle frame of the sequence). Let (u_{ij}, v_{ij}) denote the displacement of pixel (x_i, y_i) from the reference frame I to frame I_j ($i = 1..\mathcal{N}$, $j = 1..\mathcal{F}$). Let $\mathbf{U}$ and $\mathbf{V}$ denote two $\mathcal{F} \times \mathcal{N}$ matrices constructed from the displacements of all the image points across all frames:

$$\mathbf{U} = \begin{bmatrix} u_{11}, u_{21}, ..., u_{\mathcal{N}1} \\ u_{12}, u_{22}, ..., u_{\mathcal{N}2} \\ \cdot \\ \cdot \\ \cdot \\ u_{1\mathcal{F}}, u_{2\mathcal{F}}, ..., u_{\mathcal{N}\mathcal{F}} \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} v_{11}, v_{21}, ..., v_{\mathcal{N}1} \\ v_{12}, v_{22}, ..., v_{\mathcal{N}2} \\ \cdot \\ \cdot \\ \cdot \\ v_{1\mathcal{F}}, v_{2\mathcal{F}}, ..., v_{\mathcal{N}\mathcal{F}} \end{bmatrix} \tag{1}$$

Each row in these matrices corresponds to a single frame, and each column corresponds to a single point.

2.1 Ranks for Various World, Motion, and Camera Models

We next show that although the matrices $\mathbf{U}$ and $\mathbf{V}$ are large, *their ranks are very small*. In particular, we identify the ranks of the following two matrices: $\left[\frac{\mathbf{U}}{\mathbf{V}}\right]_{2\mathcal{F}\times\mathcal{N}}$ (i.e., U and V are stacked vertically), and $[\mathbf{U} \mid \mathbf{V}]_{\mathcal{F}\times 2\mathcal{N}}$ (i.e., U and V are stacked horizontally). We show that these matrices have low ranks *under many different conditions*. In the following sections we explain how to use these low-rank constraints in order to constrain the estimated flow. At no point will we need to recover any 3D quantities or camera motion. *The 3D analysis in this section is used only for deriving the upper bounds on the ranks of these matrices.*

It can be shown that the collection of all points across all views lie in a low-dimensional *variety* [19]. Under full perspective projection and discrete views, this variety is non-linear. However, there are two cases in which this variety is linear: (i) when an "affine" camera [16] is used (i.e., weak-perspective, or orthographic projection). This model is valid when the field of view is very small, and the depth fluctuations in the scene are small relative to the overall depth. (ii) when an instantaneous motion model is used (e.g., [13]). This model is valid when the camera rotation is small and the forward translation is small relative to the depth. The instantaneous model is a good approximation of the motion over *short video segments*, as the camera does not gain large motions in short periods of time. In some cases, such as airborne video, this approximation is good also for very long sequences. The instantaneous model is most relevant for this paper, as we are using short video segments for the flow analysis. Choosing the reference frame as the *middle* frame extends the applicability of the model to twice as many frames.

We have derived the linear subspace (rank) constraints for these two cases, both for a general 3D scene as well as for a planar scene. Due to lack of space, we detail the rank derivation only for one case, and provide only the final derived ranks for the other cases. The omitted derivations can be found in [12].

A 3D scene point (X_i, Y_i, Z_i) is observed at pixel (x_i, y_i) in the reference frame I. Let $\vec{t_j} = (t_{X\,j}, t_{Y\,j}, t_{Z\,j})$ denote the camera translation between frame I and frame I_j, and let $\vec{\Omega_j} = (\Omega_{X\,j}, \Omega_{Y\,j}, \Omega_{Z\,j})$ denote the camera rotation between the two frames.

I. Instantaneous motion, general 3D scene:
Under the instantaneous motion assumptions, the 2D displacement of a pixel (x_i, y_i) from I to I_j is:

$$\begin{bmatrix} u_{ij} \\ v_{ij} \end{bmatrix} = \frac{1}{Z_i} \begin{bmatrix} f t_{X\,j} - t_{Z\,j} x_i \frac{f}{f_j} \\ f t_{Y\,j} - t_{Z\,j} y_i \frac{f}{f_j} \end{bmatrix} + \begin{bmatrix} -\frac{\Omega_{X\,j}}{f_j} x_i y_i + \Omega_{Y\,j} f + \frac{\Omega_{Y\,j}}{f_j} {x_i}^2 - \Omega_{Z\,j} y_i + x_i(1 - \frac{f}{f_j}) \\ -\frac{\Omega_{X\,j}}{f_j} y_i^2 - \Omega_{X\,j} f + \frac{\Omega_{Y\,j}}{f_j} x_i y_i + \Omega_{Z\,j} x_i + y_i(1 - \frac{f}{f_j}) \end{bmatrix} \tag{2}$$

where f, f_j are the focal lengths of frames I, I_j, respectively. Eq. (2) is a straightforward rederivation of the instantaneous motion model of [13] for the case of changing focal length.

(I.a) Varying focal length (3D scene):

u_{ij} and v_{ij} can be rewritten as a *bilinear* product:

$\begin{bmatrix} u_{ij} \\ v_{ij} \end{bmatrix}_{2\times 1} = \begin{bmatrix} (M_U)_j \\ (M_V)_j \end{bmatrix}_{2\times 9} P_{i\ 9\times 1}$, where,

$P_i = (\ 1 \quad x_i \quad y_i \quad \frac{1}{Z_i} \quad \frac{x_i}{Z_i} \quad \frac{y_i}{Z_i} \quad x_i^2 \quad y_i^2 \quad (x_i y_i)\ )^T$,

is a *point-dependent* component ($i = 1..\mathcal{N}$), and

$(M_U)_j = \left(-f\Omega_{Y\,j}\ (1-\frac{f}{f_j})\ -\Omega_{Z\,j}\ f t_{X\,j}\ -\frac{f}{f_j}t_{Z\,j}\ 0\ \frac{\Omega_{Y\,j}}{f_j}\ 0\ -\frac{\Omega_{X\,j}}{f_j}\right)$

$(M_V)_j = \left(-f\Omega_{X\,j}\ \Omega_{Z\,j}\ (1-\frac{f}{f_j})\ f t_{Y\,j}\ 0\ -\frac{f}{f_j}t_{Z\,j}\ 0\ -\frac{\Omega_{X\,j}}{f_j}\ \frac{\Omega_{Y\,j}}{f_j}\right)$

are *frame-dependent* components ($j = 1..\mathcal{F}$), i.e., depends only on the camera motion and the focal length of that frame. Therefore, all flow vectors of all points across all frames can be expressed as a bilinear product of matrices:

$$\left[\frac{\mathbf{U}}{\mathbf{V}}\right]_{(2\mathcal{F}\times\mathcal{N})} = \left[\frac{\mathbf{M_U}}{\mathbf{M_V}}\right]_{(2\mathcal{F}\times 9)} \mathbf{P}_{\ (9\times\mathcal{N})} \qquad (3)$$

where the i-th column of $\mathbf{P}$ is the vector P_i, and the $j-$th row of $\mathbf{M_U}$ and $\mathbf{M_V}$ are the vectors $(M_U)_j$ and $(M_V)_j$, respectively. Therefore, $rank(\left[\frac{\mathbf{U}}{\mathbf{V}}\right]) \leq 9$.

Similarly, we can analyze the rank of $[\mathbf{U}|\mathbf{V}]$:

$(u_{ij}\ v_{ij})_{1\times 2} = M_{j\ 1\times 9}\left[(P_X)_i\ (P_Y)_i\right]_{9\times 2}$ where,

$M_j = \left(\frac{\Omega_{X\,j}}{f_j}\ \frac{\Omega_{Y\,j}}{f_j}\ f\Omega_{X\,j}\ f\Omega_{Y\,j}\ \Omega_{Z\,j}\ f t_{X\,j}\ f t_{Y\,j}\ \frac{f}{f_j}t_{Z\,j}\ (1-\frac{f}{f_j})\right)$

is a *frame-dependent* component, and

$(P_X)_i = (-x_i y_i\ \ x_i^2\ \ 0\ \ 1\ \ -y_i\ \ \frac{1}{Z_i}\ \ 0\ \ -\frac{x_i}{Z_i}\ \ x_i)^T$

$(P_Y)_i = (-y_i^2\ \ x_i y_i\ \ -1\ \ 0\ \ x_i\ \ 0\ \ \frac{1}{Z_i}\ \ -\frac{y_i}{Z_i}\ \ y_i)^T$

are *point-dependent* components. This leads to:

$$[\mathbf{U}|\mathbf{V}]_{\ (\mathcal{F}\times 2\mathcal{N})} = \mathbf{M}_{\ (\mathcal{F}\times 9)}\ [\mathbf{P_X}|\mathbf{P_Y}]_{\ (9\times 2\mathcal{N})} \qquad (4)$$

where the i-th column of $\mathbf{P_X}$ and $\mathbf{P_Y}$ are $(P_X)_i$ and $(P_Y)_i$, respectively, and the $j-$th row of $\mathbf{M}$ is M_j.

To summarize, when both the focal length and the camera motion change *across the sequence, then:*
$rank([\mathbf{U}|\mathbf{V}]) \leq 9$ *and* $rank(\left[\frac{\mathbf{U}}{\mathbf{V}}\right]) \leq 9$.

(I.b) Constant focal length (3D scene):

When the camera motion changes but the focal length remains constant across the sequence (but *not* assumed to be known), $\forall j\ f_j = f$, then the ranks of these matrices are lower [12]:
$rank([\mathbf{U}|\mathbf{V}]) \leq 6$ *and* $rank(\left[\frac{\mathbf{U}}{\mathbf{V}}\right]) \leq 8$.

II. Instantaneous motion, planar scene:

When the scene is planar, then in the perspective case [3]: $\frac{1}{Z_i} = \alpha + \beta \cdot x_i + \gamma \cdot y_i$. Substituting this expression into Eq. (2) and regrouping the terms leads to simpler bilinear forms with the following rank constraints [12]:

(I.a) Constant focal length (planar scene):

$rank(\left[\frac{\mathbf{U}}{\mathbf{V}}\right]) \leq 6$ and $rank([\mathbf{U}|\mathbf{V}]) \leq 6$

(II.b) Varying focal length (planar scene):

$rank(\left[\frac{\mathbf{U}}{\mathbf{V}}\right]) \leq 6$ and $rank([\mathbf{U}|\mathbf{V}]) \leq 8$.

III. Affine camera – 3D scene:

[18, 16] showed that in the case of an affine camera, the corresponding image points across all image frames lie in a 4-dimensional linear subspace (and with some additional manipulation, it can be reduced to 3). The derivation of subspace constraints for optical flow is very similar, leading to the following rank constraints:
$rank(\left[\frac{\mathbf{U}}{\mathbf{V}}\right]) \leq 4$ *and* $rank([\mathbf{U}|\mathbf{V}]) \leq 8$.

IV. Affine camera, planar scene:

$rank(\left[\frac{\mathbf{U}}{\mathbf{V}}\right]) \leq 3$ *and* $rank([\mathbf{U}|\mathbf{V}]) \leq 6$.

Remarks: We showed that when the camera motion changes across the sequence (and possibly also the focal length), then the ranks of these matrices for a wide variety of models are all within a small range (≤ 9), and are significantly lower than the actual size of these matrices ($\mathcal{F}\times 2\mathcal{N}$ and $2\mathcal{F}\times\mathcal{N}$). We will use these rank constraints alone to constrain the flow estimation. *No* 3D information will be recovered in this process. Furthermore, the *actual rank* of these matrices may be even *lower* than the derived theoretical upper bounds, e.g., in cases when the camera motion is degenerate (e.g., uniform) across the sequence. As will be explained in Sect. 4.3, our algorithm *automatically* detects the actual underlying ranks, directly from image brightness quantities, *prior* to computing the flow. This implies that the rank constraint can be applied to a sequence of frames *without the need to a-priori determine the underlying model, or its degeneracies.*

3 Subspace Constraints on Image Brightness

The straightforward way to take advantage of the subspace constraints is to first compute inter-frame flow fields using an existing two-frame flow estimation technique, and then project the collection of all these flow fields into the appropriate lower dimensional subspace. However, there are two problems with this two-stage approach: (i) all flow-vectors are treated equally, without regard to their reliability, and (ii) the flow-fields resulting from the unconstrained two-frame flow estimation (first step) may contain flow-vectors which are so erroneous, that the subspace projection will not suffice to correct them. Moreover, if a significant number of flow vectors is severely corrupted, these may severely damage all other flow-vectors.

To avoid these two problems, we propose a one-stage approach for applying the low-dimensionality subspace constraints directly to *measurable image quantities*

even during the flow estimation process itself. This approach implicitly leads to *confidence-weighted subspace projection* of the data, in accordance with the amount of local image structure at each pixel. In particular, we derive two different brightness subspace constraints: (i) a multi-point multi-frame *point-based* constraint, which is based on the *brightness constancy equation* (Sect. 3.1), and (ii) a multi-point multi-frame *region-based* constraint, which is based on the *Lucas & Kanade formulation* (Sect. 3.2). The benefits of using these constraints is explained in Sect. 4.

3.1 The Generalized Brightness Constancy Constraint

Let (x_i, y_i) be a pixel in the reference frame I, whose corresponding pixel in another frame I_j is (x_i+u_{ij}, y_i+v_{ij}). The Brightness Constancy Equation, which is defined on a single pixel between two frames, states that: $I_j(x_i, y_i) = I(x_i - u_{ij}, y_i - v_{ij})$. For *very small* (u_{ij}, v_{ij}), this equation can be linearized as: $u_{ij} \cdot I_{x_i} + v_{ij} \cdot I_{y_i} I_{y_i} + I_{tij} = 0$, where I_{x_i}, I_{y_i} are the spatial derivatives of the reference frame I at pixel (x_i, y_i), and I_{tij} is the temporal derivative: $I_{tij} = (I_j(x_i, y_i) - I(x_i, y_i))$.

However, in practice, (u_{ij}, v_{ij}) may not be small, especially when dealing with multiple frames. To increase its range of applicability to larger (u_{ij}, v_{ij}), the linearization can be applied within an iterative (coarse-to-fine) refinement process [3]. Let (u^0_{ij}, v^0_{ij}) be the current estimate of (u_{ij}, v_{ij}) during an iterative estimation process. Let $\Delta u_{ij} = u_{ij} - u^0_{ij}$ and $\Delta v_{ij} = v_{ij} - v^0_{ij}$. The Brightness Constancy Equation can be rewritten as: $I_j(x_i + u^0_{ij}, y_i + v^0_{ij}) = I(x_i - \Delta u_{ij}, y_i - \Delta v_{ij})$. Assuming small $(\Delta u_{ij}, \Delta v_{ij})$, this equation can be linearized as:

$\Delta u_{ij} I_{x_i} + \Delta v_{ij} I_{y_i} + (I_j(x_i+u^0_{ij}, y_i+v^0_{ij}) - I(x_i, y_i)) = 0$

Because the subspace constraints are defined on the displacements (u_{ij}, v_{ij}) and not on the increments (see Sect. 2), we substitute the expression for $(\Delta u_{ij}, \Delta v_{ij})$, leading to the following form of the brightness constancy equation, which we will use:

$$u_{ij} \cdot I_{x_i} + v_{ij} \cdot I_{y_i} \;=\; -I^0_{t\,ij}, \qquad (5)$$

where,

$I^0_{t\,ij} = (I_j(x_i+u^0_{ij}, y_i+v^0_{ij}) - I(x_i, y_i) - u^0_{ij} I_{x_i} - v^0_{ij} I_{y_i})$.

Eq. (5) provides a single *line constraint* on the two unknowns u_{ij}, v_{ij}, and hence does not suffice for uniquely determining the unknown displacement of a single pixel between two frames.

Let $I_1, ..., I_{\mathcal{F}}$ be a sequence of frames, as defined in Sect. 2. The collection of all Brightness Constancy Constraints (Eq. (5)) of all image points across all image frames can be compactly written in a single *matrix form* as:

$$\underset{(\mathcal{F}\times 2\mathcal{N})}{\left[\;\mathbf{U}\;\middle|\;\mathbf{V}\;\right]} \cdot \underset{(2\mathcal{N}\times\mathcal{N})}{\left[\begin{array}{c}\mathbf{F_X}\\ \hline \mathbf{F_Y}\end{array}\right]} = \underset{(\mathcal{F}\times\mathcal{N})}{\mathbf{F_T}} \qquad (6)$$

where $\mathbf{F_X}$ and $\mathbf{F_Y}$ are $\mathcal{N}\times\mathcal{N}$ diagonal matrices with the spatial x- and y- derivatives of the reference frame I in their diagonal:

$$\mathbf{F_X} = \begin{bmatrix} I_{x_1} & 0 & \dots & 0 \\ 0 & I_{x_2} & \dots & 0 \\ & & \cdot & \\ & & \cdot & \\ & & \cdot & \\ 0 & 0 & \dots & I_{x_\mathcal{N}} \end{bmatrix} \quad \mathbf{F_Y} = \begin{bmatrix} I_{y_1} & 0 & \dots & 0 \\ 0 & I_{y_2} & \dots & 0 \\ & & \cdot & \\ & & \cdot & \\ & & \cdot & \\ 0 & 0 & \dots & I_{y_\mathcal{N}} \end{bmatrix}$$

and $\mathbf{F_T}$ is an $\mathcal{F}\times\mathcal{N}$ matrix of the *temporal* derivatives (of all image points across all frames) estimated at the current stage of the iterative process, namely:

$$\mathbf{F_T} = \begin{bmatrix} -I^0_{t\,11} & -I^0_{t\,21} & \dots & -I^0_{t\,\mathcal{N}1} \\ -I^0_{t\,12} & -I^0_{t\,22} & \dots & -I^0_{t\,\mathcal{N}2} \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ -I^0_{t\,1\mathcal{F}} & -I^0_{t\,2\mathcal{F}} & \dots & -I^0_{t\,\mathcal{N}\mathcal{F}} \end{bmatrix}$$

The matrices $\mathbf{F_X}$, $\mathbf{F_Y}$, and $\mathbf{F_T}$, contain only *measurable image quantities*. The matrices $\mathbf{U}$ and $\mathbf{V}$ contain all the *unknown displacements*. Note that all flow-vectors corresponding to a single scene point share the same *spatial* derivatives I_{x_i}, I_{y_i} (as these are computed in the reference frame I, and are independent of the other frame j). However, their *temporal* derivatives $I_{t_{ij}}$ *do* vary from frame to frame (and in every iteration). We refer to the multi-point multi-frame Eq. (6) as the *the Generalized Brightness Constancy Equation.*

Note that when no additional information on $[\mathbf{U}|\mathbf{V}]$ is used, then Eq. (6) is no more than the collection of all the individual two-frame brightness constancy equations of Eq. (5). However, this matrix formulation allows us to apply rank constraints directly to measurable image quantities. For example, $rank([\mathbf{U}|\mathbf{V}]) \leq r$ implies that $rank(\mathbf{F_T}) \leq r$. We can therefore apply the rank constraint directly to the data matrix $\mathbf{F_T}$ *prior* to solving for the displacements $\mathbf{U}$ and $\mathbf{V}$. This formulation, as well as the one which is next described in Sect. 3.2, form the basis for our direct multi-point multi-frame algorithm, which is described in Sect. 4.

3.2 The Generalized Lucas & Kanade Constraint

Lucas and Kanade [14] extended the pixel-based brightness constancy constraints of Eq. (5) to a local region-based constraint, by assuming a uniform displacement in very small windows (typically 3×3 or 5×5). Then, for each pixel (x_i, y_i), they solve for its displacement vector (u_{ij}, v_{ij}) by minimizing the following local error measure $E(u_{ij}, v_{ij})$ within its neighborhood (window) W_i:

$$E(u_{ij}, v_{ij}) = \sum_{k\in W_i} (u_{ij} \cdot I_{x_k} + v_{ij} \cdot I_{y_k} + I^0_{t\,kj})^2$$

(The Lucas and Kanade equation was slightly modified to fit our iterative notation). Differentiating the error $E(u_{ij}, v_{ij})$ with respect to u_{ij} and v_{ij}, and setting these derivatives to zero, yields a set of two linear equations in the two unknown displacement components (u_{ij}, v_{ij}) for each pixel:

$$[u_{ij}\ v_{ij}]_{1\times 2} \cdot \left[\begin{array}{cc} a_i & b_i \\ b_i & c_i \end{array}\right]_{\mathbf{2\times 2}} = [g_{ij} h_{ij}]_{1\times 2} \qquad (7)$$

$a_i, b_i, c_i, g_{ij}, h_{ij}$ are measurable image quantities:
$a_i = \sum_k (I_{x_k})^2$, $\quad b_i = \sum_k (I_{x_k} \cdot I_{y_k})$, $\quad c_i = \sum_k (I_{y_k})^2$,
$g_{ij} = -\sum_k (I_{x_k} \cdot {I^0}_{t_{kj}})$, $\quad h_{ij} = -\sum_k (I_{y_k} \cdot {I^0}_{t_{kj}})$.
a_i, b_i, c_i are computed in the reference image I, and are independent of j. g_{ij}, h_{ij} depend on both.
Eq. (7) provides *two* equations on the two unknowns u_{ij}, v_{ij}, as opposed to Eq. (5), which provides only one. This is because of the uniform-displacement assumption within the local windows. While this assumption imposes a type of *local* smoothness constraint, it only affects the accuracy of the flow estimation within the small window, but does not propagate these errors to other image regions (as opposed to *global* smoothness (e.g., [10])). The vector (u_{ij}, v_{ij}) therefore has a unique solution when the coefficient matrix $\left[\begin{array}{cc} a_i & b_i \\ b_i & c_i \end{array}\right]$ is not singular (e.g., for corners and textured areas). For image regions, where the local information is insufficient (e.g., edges), the matrix will be singular. In these regions the flow vector (u_{ij}, v_{ij}) cannot be uniquely determined even by the Lucas & Kanade algorithm. Under Gaussian noise assumptions, the matrix $\left[\begin{array}{cc} a_i & b_i \\ b_i & c_i \end{array}\right]$ in Eq. (7) can be shown to be the *posterior inverse covariance matrix* of the estimated flow vector (u_{ij}, v_{ij}).

Now, considering multiple-points over multiple-frames. As in the case of the Generalized Brightness Constancy Equation (6), all the flow-vectors (u_{ij}, v_{ij}) from a reference pixel (x_i, y_i) in I to all other frames I_j $(j = 1..\mathcal{F})$ share the *same coefficient (inverse covariance) matrix* $\left[\begin{array}{cc} a_i & b_i \\ b_i & c_i \end{array}\right]$ in their two-frame Lucas & Kanade constraints (Eq. (7)). Hence, all the Lucas & Kanade constraints on *all points* $(i = 1..\mathcal{N})$ across *all frames* $(j = 1..\mathcal{F})$ can be compactly written in a single matrix form as:

$$\left[\begin{array}{c|c} \mathbf{U} & \mathbf{V} \end{array}\right]_{(\mathcal{F}\times 2\mathcal{N})} \cdot \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{B} & \mathbf{C} \end{array}\right]_{(2\mathcal{N}\times 2\mathcal{N})} = \left[\begin{array}{c|c} \mathbf{G} & \mathbf{H} \end{array}\right]_{(\mathcal{F}\times 2\mathcal{N})} \qquad (8)$$

where $\mathbf{U}$ and $\mathbf{V}$ are as defined in Eq. (1). The three $\mathcal{N}\times\mathcal{N}$ diagonal matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are constructed from the coefficient values a_i, b_i, c_i, respectively:

$$\mathbf{A} = \left[\begin{array}{cccc} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ 0 & 0 & \dots & a_{\mathcal{N}} \end{array}\right] \qquad \mathbf{B} = \left[\begin{array}{cccc} b_1 & 0 & \dots & 0 \\ 0 & b_2 & \dots & 0 \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ 0 & 0 & \dots & b_{\mathcal{N}} \end{array}\right]$$

$$\mathbf{C} = \left[\begin{array}{cccc} c_1 & 0 & \dots & 0 \\ 0 & c_2 & \dots & 0 \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ 0 & 0 & \dots & c_{\mathcal{N}} \end{array}\right]$$

The two $\mathcal{F}\times\mathcal{N}$ matrices $\mathbf{G}$ and $\mathbf{H}$ are constructed from the values g_{ij}, h_{ij}:

$$\mathbf{G} = \left[\begin{array}{cccc} g_{11} & g_{21} & \dots & g_{\mathcal{N}1} \\ g_{12} & g_{22} & \dots & g_{\mathcal{N}2} \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ g_{1\mathcal{F}} & g_{2\mathcal{F}} & \dots & g_{\mathcal{N}\mathcal{F}} \end{array}\right] \qquad \mathbf{H} = \left[\begin{array}{cccc} h_{11} & h_{21} & \dots & h_{\mathcal{N}1} \\ h_{12} & h_{22} & \dots & h_{\mathcal{N}2} \\ & \cdot & & \\ & \cdot & & \\ & \cdot & & \\ h_{1\mathcal{F}} & h_{2\mathcal{F}} & \dots & h_{\mathcal{N}\mathcal{F}} \end{array}\right]$$

We refer to the multi-point multi-frame Eq. (8) as the *the Generalized Lucas & Kanade Equation.*

When no additional information on $[\mathbf{U}|\mathbf{V}]$ is used, then Eq. (8) is no more than the collection of all the individual two-frame equations of Eq. (7). However, as before, if we know that $rank([\mathbf{U}|\mathbf{V}]) \leq r$, it entails that $rank([\mathbf{G}\,|\,\mathbf{H}]) \leq r$. Since $[\mathbf{G}\,|\,\mathbf{H}]$ is a matrix constructed from *known measurable image quantities*, applying the rank constraint to it *prior* to solving for $[\mathbf{U}\,|\,\mathbf{V}]$ will constrain the flow estimation process itself. The interpretation of this operation is explained below.

Confidence Weighted Subspace Projection:
Note that applying the rank constraint to $[\mathbf{G}\,|\,\mathbf{H}]$ is in fact equivalent to applying the rank constraint directly to the flow-vector matrix $[\mathbf{U}\,|\,\mathbf{V}]$, but after first weighting the individual flow vectors (u_{ij}, v_{ij}) with their corresponding individual *inverse covariance matrices* $\left[\begin{array}{cc} a_i & b_i \\ b_i & c_i \end{array}\right]$. This means that more reliable flow-vectors will have more influence in the subspace projection process, while less reliable vectors will have smaller influence. Applying the rank constraint to $[\mathbf{G}\,|\,\mathbf{H}]$ therefore has the effect of *confidence-weighted subspace projection* of all the flow-vectors *prior* to computing them. This is used to constrain the flow estimation process itself in Sect. 4.

4 Multi-Frame Multi-Point Algorithm

Let r_1 and r_2 denote the ranks of $[\mathbf{U}|\mathbf{V}]$ and $\left[\frac{\mathbf{U}}{\mathbf{V}}\right]$, respectively. We utilize the "brightness subspace constraints" of Eqs. (6) and (8) in two ways:

4.1 Noise Reduction in Image Measurements:

The measurement matrices $\mathbf{F_T}$ and $[\mathbf{G}|\mathbf{H}]$ are projected onto lower-rank matrices $\hat{\mathbf{F}}_\mathbf{T}$ and $\left[\hat{\mathbf{G}}|\hat{\mathbf{H}}\right]$ of rank r_1. We know that $r_1 \leq 9$ (see Sect. 2.1), but in practice, the actual rank of these matrices may be even lower than the theoretical upper bound of 9. The actual rank can be automatically detected from these measurement matrices, as explained in Sect. 4.3.

The rank-reduction process inhibits noisy measurements in the measurement matrices. It can be directly

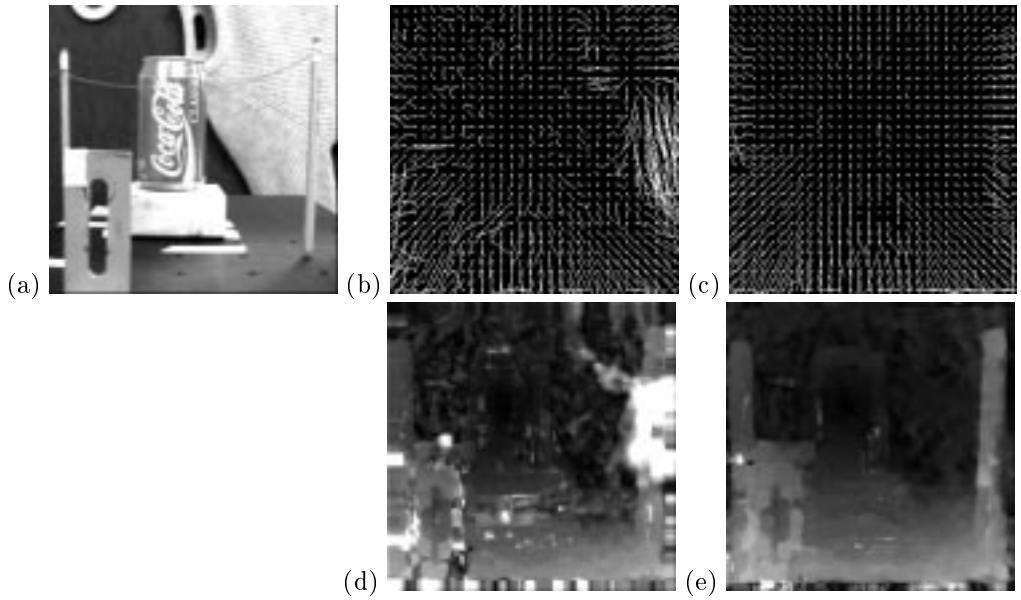

Figure 1: *Real image sequence (the NASA coke-can sequence).* **(a)** One frame from a 27-frame sequence of a forward moving camera in a 3D scene. **(b)** Flow field generated with the two-frame Lucas & Kanade algorithm. Note the errors in the right hand side, where there is depth discontinuity (pole in front of sweater), as well as the aperture problem. **(c)** The flow field for the corresponding frame generated by the multi-frame constrained algorithm. Note the good recovery of flow in those regions. **(d,e)** The flow magnitudes at every pixel. This display provides a higher resolution display of the error. Note the clear depth discontinuities in the multi-frame flow image. The flow values on the coke can are very small, because the camera FOE is in that area.

applied to [**G**|**H**]. Alternatively, since temporal derivatives $I_{t_{ij}}$ are typically the most noisy image measurements (because of misalignment errors and subpixel interpolation), the rank reduction can be *first* applied to $\mathbf{F_T}$. This step gives more accurate temporal derivatives. These noise-reduced temporal derivatives can then be used to compute [**G**|**H**] using Eq. (7). [**G**|**H**] is then further projected onto a lower-rank matrix $\left[\hat{\mathbf{G}}|\hat{\mathbf{H}}\right]$. This corresponds to applying *confidence-weighted subspace projection* on the flow vectors *prior* to computing them (see Sect. 3.2).

Now that *local* noisy measurements have been inhibited via the *global* subspace constraints, we proceed to computing an initial estimate $[\mathbf{U_0}\,|\,\mathbf{V_0}]$ for all flow vectors by solving: $\left[\begin{array}{c|c}\mathbf{U_0}&\mathbf{V_0}\end{array}\right] = \left[\begin{array}{c|c}\hat{\mathbf{G}}&\hat{\mathbf{H}}\end{array}\right] \cdot \left[\begin{array}{c|c}\mathbf{A}&\mathbf{B}\\\hline\mathbf{B}&\mathbf{C}\end{array}\right]^+$ (where M^+ denotes the *pseudo-inverse* of a matrix M). Note that because of the diagonal structure of $\mathbf{A}, \mathbf{B}, \mathbf{C}$, the matrix $\left[\begin{array}{c|c}\mathbf{A}&\mathbf{B}\\\hline\mathbf{B}&\mathbf{C}\end{array}\right]^+$ consists of the individual pseudo-inverse matrices $\left[\begin{array}{cc}a_i&b_i\\b_i&c_i\end{array}\right]^+$. This step therefore yields accurate flow for pixels with enough local image structure (i.e., pixels whose inverse covariance matrix is non-singular). For other pixels, it accurately estimates only the component of the flow in the direction of the gradient, which is the *normal flow* (because pseudo-inverse estimation yields the solution with smallest norm). This is addressed next.

4.2 Eliminating the Aperture Problem

We use the rank constraint on $\left[\frac{\mathbf{U}}{\mathbf{V}}\right]$ to determine the missing components of flow vectors at pixels with insufficient local image structure. $rank(\left[\frac{\mathbf{U}}{\mathbf{V}}\right]) = r_2$ implies that there is a decomposition:

$$\underset{(2\mathcal{F}\times\mathcal{N})}{\left[\frac{\mathbf{U}}{\mathbf{V}}\right]} = \underset{(2\mathcal{F}\times r_2)}{\mathbf{K}} \cdot \underset{(r_2\times\mathcal{N})}{\mathbf{L}} = \left[\frac{\mathbf{K_U}}{\mathbf{K_V}}\right] \cdot \mathbf{L} \qquad (9)$$

where $\mathbf{K_U}$ and $\mathbf{K_V}$ are the upper and lower halves of the matrix $\mathbf{K}$. The columns of $\mathbf{K}$ form a basis which spans the subspace of all columns of $\left[\frac{\mathbf{U}}{\mathbf{V}}\right]$. The columns of $\mathbf{L}$ are the coefficients in the linear combination. This decomposition is of course not unique. However, if there are more than r_2 pixels whose correspondences across all frames can be reliably computed, then these flow vectors could be used to generate a basis $\mathbf{K}$. The $\left[\frac{\mathbf{U_0}}{\mathbf{V_0}}\right]$ computed in the previous step, give accurate flow vectors for pixels whose local inverse covariance matrix $\left[\begin{array}{cc}a_i&b_i\\b_i&c_i\end{array}\right]$ is well conditioned. These flow vectors are used to generate a basis $\mathbf{K}$. Once a basis has been computed, the number of unknowns shrink from the original number of $2\mathcal{FN}$ unknown unconstrained displacements to $\mathcal{N}r_2$ unknowns, which are

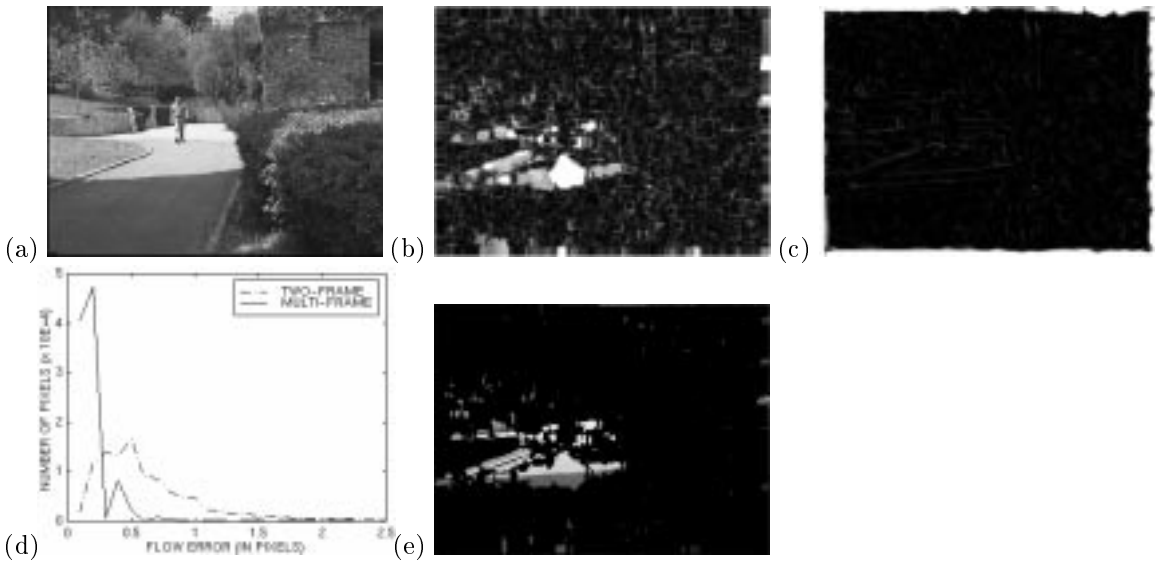

Figure 2: *Synthetic sequence with ground truth – a quantitative comparison.* **(a)** One out of a 10-frame sequence. The sequence was synthetically generated by applying a set of 3-D consistent homographies to warp a single image. This provides ground truth on the flow. **(b,c)** Error maps showing magnitudes of errors between the ground truth flow and the computed flow field. (b) shows errors for the two-frame Lucas & Kanade algorithm. (c) shows errors for the multi-frame constrained algorithm for the corresponding frame. Brighter values correspond to larger errors. **(d)** A histogram of the errors in both flow fields. Flow values at image boarders were ignored. In the multi-frame method almost all errors are smaller than 0.2 pixel, and all are smaller than 0.5 pixel. In the two-frame method, most flow vectors have an error of *at least* 0.5 pixel. **(e)** The image regions for which the errors in the *two-frame* method exceeded 1.0 pixel. These, as expected, correspond to areas which suffer from the aperture problem. The subspace constrained algorithm accurately recovered the flow even in those regions.

the unknown components of $\mathbf{L}$. Note that both $\mathbf{U}$ and $\mathbf{V}$ share the *same* coefficients $\mathbf{L}$. Hence, for flow vectors with only one known flow-component (e.g., the normal-flow), the other component can be uniquely determined via this decomposition (which is not true in the equivalent decomposition of $[\mathbf{U}|\mathbf{V}]$). Plugging the decomposition of Eq. (9) into Eq. (6) leads to a set of $\mathcal{FN}$ *linear* equations in the $\mathcal{N}r_2$ unknowns:

$$[\mathbf{K}_{\mathrm{U}}\mathbf{L} \,|\, \mathbf{K}_{\mathrm{V}}\mathbf{L}] \cdot \left[\frac{\mathbf{F_X}}{\mathbf{F_Y}}\right] = \hat{\mathbf{F}}_{\mathbf{T}}. \qquad (10)$$

This set of equations is overconstrained if the number of frames $\mathcal{F}$ is larger than r_2 (where r_2 is the lowest of the *actual* rank and the theoretical upper bound).

Similarly, plugging the decomposition of Eq. (9) into Eq. (8) leads to an alternative set of linear equations, with *twice* as many equations ($2\mathcal{FN}$ equations) in the same $\mathcal{N}r_2$ unknowns:

$$[\mathbf{K}_{\mathrm{U}}\mathbf{L} \,|\, \mathbf{K}_{\mathrm{V}}\mathbf{L}] \cdot \left[\begin{array}{c|c}\mathbf{A} & \mathbf{B}\\ \hline \mathbf{B} & \mathbf{C}\end{array}\right] = \left[\begin{array}{c|c}\hat{\mathbf{G}} & \hat{\mathbf{H}}\end{array}\right]. \qquad (11)$$

This set of equations is thus overconstrained if the number of frames $\mathcal{F}$ is larger than $\frac{1}{2}r_2$. Each of the two abovementioned options has its advantages: Eq. (11) is numerically more stable (because of the local confidence-weighted averaging over the small (3×3 or 5×5) windows from the Lucas & Kanade algorithm, and because there are twice as many equations), but this benefit comes with the price of lower spatial resolution in the flow recovery. On the other hand, Eq. (10) provides half as many equations, but allows for *higher spatial resolution* of flow recovery, as it does not use the small window averaging. In the current implementation of our algorithm we used Eq. (11). We now summarize the algorithm.

4.3 The multi-point multi-frame algorithm:

1. Construct a Gaussian pyramid for all image frames.

2. For each iteration in each pyramid level do:

(a) Compute matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{G}, \mathbf{H}$.

(b) Project $[\mathbf{G} \,|\, \mathbf{H}]$ onto lower-rank (r_1) matrix $\left[\hat{\mathbf{G}} \,\middle|\, \hat{\mathbf{H}}\right]$.

(c) Compute an initial flow estimate $[\mathbf{U_0} \,|\, \mathbf{V_0}]$:

$$\left[\begin{array}{c|c}\mathbf{U_0} & \mathbf{V_0}\end{array}\right] = \left[\begin{array}{c|c}\hat{\mathbf{G}} & \hat{\mathbf{H}}\end{array}\right] \cdot \left[\begin{array}{c|c}\mathbf{A} & \mathbf{B}\\ \hline \mathbf{B} & \mathbf{C}\end{array}\right]^{+}.$$

(d) Compute an r_2-dimensional basis $\mathbf{K}$ from the columns of $\left[\frac{\mathbf{U_0}}{\mathbf{V_0}}\right]$

(e) Linearly solve for the unknown matrix $\mathbf{L}$ using either Eq. (10) (*Generalized Brightness Constancy*) or (11) (*Generalized Lucas&Kanade*). This step recovers the missing components of normal-flow vectors and produces more accurate flow estimates $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$.

3. Keep iterating to refine $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$.

Step (b) reduces noise in the measurements, while steps (d) and (e) eliminate the aperture problem. *When the algorithm is applied to two frames, and steps (b),(d),(e) are skipped, it reduces to an iterative coarse-to-fine version of the Lucas&Kanade algorithm [3].* Step (a) can be preceded by projecting the matrix $\mathbf{F_T}$ onto a lower-rank matrix $\hat{\mathbf{F}}_\mathbf{T}$, as discussed in Sect. 4.1. This step is not yet incorporated in our current implementation (hence omitted from the algorithm), but is expected to further reduce the noise in the measurement matrix $[\mathbf{G} \mid \mathbf{H}]$ prior to its own rank-reduction.

Automatic Rank Detection:

Step (b) projects matrices onto lower-rank matrices, as defined in Sect. 2.1. In practice, the actual rank of these matrices, with some allowed noise tolerance, may be even lower than the theoretical upper bound r_1 (e.g., in cases of degenerate camera motions or scene structures). We automatically detect the *actual* rank of these matrices: Let $\mathbf{M}$ be a $k \times l$ matrix, with a known upper bound r on its rank, and an *actual* rank r_M ($r_M \leq r$). The rank reduction (i.e., subspace projection) of $\mathbf{M}$ is done by applying Singular Value Decomposition to $\mathbf{M}$. We check for the existence of a lower rank $r' < r$ such that $(\sum_{i=r'+1}^{m} \lambda_i^2)/(\sum_{i=1}^{m} \lambda_i^2) < \epsilon$, where m is the number of eigenvalues: $m = min(k, l)$, and ϵ allows for some noise tolerance (we use $\epsilon = 1\%$). r_M is set to be $min(r, r')$. All singular values other than the r_M largest ones are then set to zero, and the matrices produced in the SVD step are re-composed, yielding a matrix $\hat{\mathbf{M}}$ of rank r_M (which is closest to $\mathbf{M}$ in the Frobenius norm). Step (d) uses the same SVD procedure to estimate a spanning basis $\mathbf{K}$.

Results:

Figs. 1 and 2 show comparisons of the multi-frame constrained algorithm with an iterative coarse-to-fine version of the two-frame Lucas & Kanade algorithm. The latter is computed by using our multi-frame algorithm (see Sect. 4.3), but without applying the subspace projection steps (b),(d),(e). This allows us to *isolate* the effects of subspace projection on the accuracy of the flow estimation. The comparison is done both for real data, as well as for synthetic data with ground truth. For further details regarding the experiments and the results, see figure captions.

Acknowledgement

The author would like to thank P. Anandan for the helpful discussions and for his insightful comments about the topic and the paper. Thanks are also due to R. Szeliski, R. Basri, and L. Zelnik-Manor for their useful comments on the paper.

References

[1] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *IJCV*, 2:283–310, 1989.

[2] J.L. Barron, D.J. Fleet, S.S. Beauchemin, and T.A. Burkitt. Performance of optical flow techniques. In *CVPR*, pages 236–242, Champaign, June 1992.

[3] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *ECCV*, pages 237–252, May 1992.

[4] J.R. Bergen, P.J. Burt, R. Hingorani, and S. Peleg. A three-frame algorithm for estimating two-component image motion. *PAMI*, 14:886–895, September 1992.

[5] M.J. Black and P. Anandan. Robust dynamic motion estimation over time. In *CVPR*, pages 296–302, 1991.

[6] M.J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 63:75–104, 1996.

[7] K. Hanna. Direct multi-resolution estimation of ego-motion and structure from motion. In *IEEE Workshop on Visual Motion*, pp. 156–162, Princeton, 1991.

[8] K. J. Hanna and N. E. Okamoto, Combining Stereo and Motion for Direct Estimation of Scene Structure. In *ICCV*, pages 357-365, 1993.

[9] D.J. Heeger and A.D. Jepson. Subspace methods for recovering rigid motion I: Algorithm and implementation. *IJCV*, 7:95–117, 1992.

[10] B.K.P. Horn and B.G. Schunck. Determining optical flow. *AI*, 17(1–3):185–203, August 1981.

[11] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *IJCV*, 12:5–16, 1994.

[12] M. Irani. Multi-Frame Correspondence Estimation Using Subspace Constraints. *TR in preparation*, 1999.

[13] 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.

[14] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IUW*, pages 121–130, 1981.

[15] H. H. Nagel. Displacement vectors derived from second order intensity variations in intensity sequences. *Computer Vision, Pattern recognition and Image Processing*, 21:85–117, 1983.

[16] L.S. Shapiro. *Affine Analysis of Image Sequences.* Cambridge University Press, Cambridge, UK, 1995.

[17] G.P. Stein and A. Shashua. Model-based brightness constraints: On direct estimation of structure and motion. In *CVPR*, pages 400–406, June 1997.

[18] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *IJCV*, 9:137–154, 1992.

[19] P.H.S. Torr. Geometric motion segmentation and model selection. *Proceedings of The Royal Society of London A*, 356:1321–1340, 1998.