Décomposition de déformation pour l'estimation d'un mouvement de caméra

Claire Jonchery[1], Françoise Dibos[1], Georges Koepfler[2]

[1]Laboratoire CEREMADE, Université Paris Dauphine
Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France

[2]Laboratoire MAP5, Université Paris V
45, rue des Saints-Pères, 5270 Paris Cedex 06, France
jonchery@ceremade.dauphine.fr,dibos@ceremade.dauphine.fr
Georges.Koepfler@math-info.univ-paris5.fr

Résumé − Nous proposons dans ce papier une nouvelle approche globale pour estimer le mouvement d'une caméra filmant une scène fixe. Notre méthode considère deux images consécutives d'une séquence et utilise le contenu de ces images et le flot optique sur la séquence. L'originalité de notre approche repose sur une nouvelle modélisation du mouvement de caméra permettant de séparer la déformation entre deux images consécutives en deux composantes : une similitude et une déformation "purement" projective. La similitude est, dans un premier temps, estimée à partir des deux images par un algorithme d'estimation de mouvement paramétrique 2D. Dans un second temps, la déformation projective est évaluée à partir du flot optique et d'un critère de recalage entre les images.

Abstract − In this paper, we propose a new global approach for estimating motion parameters of a camera which films a static scene. Our method deals with adjacent frames of a sequence and utilizes both images content and optical flow. Thanks to an original model of camera motion, our approach separates the deformation between two adjacent frames in two components : a similarity and a "purely" projective application. The similarity is, in a first time, estimated from the two images with a parametric 2D motion model and then, the projective deformation is evaluated from the optical flow and from a registration criterion.

1 Introduction

L'estimation du mouvement d'une caméra est un problème difficile car le mouvement d'un pixel entre deux images dépend non seulement des six paramètres du mouvement de la caméra mais aussi de la profondeur du point de la scène projeté. La littérature présente un grand nombre de modèles différents et de techniques d'estimation variées, séparant ou non l'estimation du mouvement de celle de la structure de la scène. On distingue classiquement trois types d'approches. Les méthodes discrètes d'abord, utilisent des correspondances de points entre les images ; dans [2] par exemple, les appariements mènent à une estimation de la matrice essentielle permettant de déduire le mouvement de la caméra. Les méthodes différentielles ensuite, utilisent le flot optique préalablement calculé entre les images ; un grand nombre de ces méthodes est basé sur la contrainte épipolaire différentielle [6], sur la contrainte bilinéaire [3] ou sur le mouvement de parallaxe [7]. Une présentation et une comparaison de plusieurs techniques, dont la donnée de base est le flot optique, est présentée dans [8]. Les méthodes directes enfin, utilisent directement le contenu des images données. Elles reposent sur la contrainte d'illumination constante, à laquelle on ajoute une hypothèse sur le mouvement et sur les profondeurs de la scène filmée [4].

Notre méthode d'estimation se situe dans le contexte particulier, présenté dans la section 2, d'estimation du mouvement d'une caméra entre deux images consécutives dans une séquence. Dans ce cadre, nous approximons la déformation produite entre les deux images par une déformation plane, elle-même décomposée en une similitude et une application "purement" projective dans la section 3. Grâce à une approximation du flot optique associée à cette décomposition, nous élaborons un algorithme d'estimation du mouvement de caméra, proposé dans la section 4. La section 5 présente plusieurs résultats.

2 Contexte et modélisation

Considérons deux images f et g consécutives dans une séquence. Même dans le cas où la vitesse de la caméra est conséquente, le nombre élevé d'images par seconde limite nécessairement le mouvement de la caméra entre deux acquisitions consécutives. Les deux images f et g sont donc très proches l'une de l'autre et l'effet de parallaxe, n'affectant qu'un très petit nombre de pixels, peut être négligé. On peut aller plus loin ; si la scène filmée est suffisamment éloignée du centre optique, on peut la supposer plane et orthogonale à la direction de l'axe de la caméra avant le déplacement : la structure de la scène est négligée et l'image g est considérée comme une déformation plane de l'image f. Le problème d'estimation du mouvement de

caméra devient alors un problème d'estimation de transformations planes. Notons que plus la scène filmée est éloignée de la caméra, plus cette hypothèse est réaliste.

Nous utilisons le modèle de caméra sténopé classique. Soit (C, i, j, k) le repère orthonormal associé à la caméra, C représentant la position du centre optique et k la direction de l'axe optique. La projection m d'un point M, de coordonnées (X, Y, Z) dans le repère (C, i, j, k), est l'intersection du rayon optique (CM) avec le plan image $\mathcal{R} : \{Z = f_c\}$. Si c est l'intersection de l'axe optique avec $\mathcal{R}$, les coordonnées (x, y) de m dans le repère (c, i, j) du plan $\mathcal{R}$, sont données par

$$\begin{cases} x = f_c \dfrac{X}{Z} \\ \\ y = f_c \dfrac{Y}{Z}. \end{cases}$$

Comme la longueur focale f_c agit comme un facteur d'échelle sur l'image, on peut choisir en toute généralité de fixer f_c à 1.

Considérons maintenant un mouvement de caméra $D = (R, T)$ où R est une rotation d'axe contenant C et T une translation de vecteur t. Le repère associé à la caméra (C, i, j, k) est ainsi transformé par le déplacement D en $(C', R(i), R(j), R(k))$, avec $\overrightarrow{CC'} = t$. Dans ce cadre, en notant

$$R = \begin{pmatrix} a & b & c \\ a' & b' & c' \\ a'' & b'' & c'' \end{pmatrix} \text{ et } t = \alpha i + \beta j + \gamma k,$$

les images f et g sont liées par la relation

$$\begin{aligned} g(x, y) &= (\varphi f)(x, y) = f(\varphi(x, y)) \\ &= f\left(\frac{ax + by + c + \widetilde{\alpha}}{a''x + b''y + c'' + \widetilde{\gamma}}, \frac{a'x + b'y + c' + \widetilde{\beta}}{a''x + b''y + c'' + \widetilde{\gamma}} \right) \end{aligned}$$

où $\widetilde{t} = (\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\gamma})$ est le vecteur de translation quotienté par la profondeur Z_0 de la scène filmée. La matrice $3D$ associée à φ peut s'écrire

$$\mathcal{M}_\varphi = \begin{pmatrix} a & b & c + \widetilde{\alpha} \\ a' & b' & c' + \widetilde{\beta} \\ a'' & b'' & c'' + \widetilde{\gamma} \end{pmatrix} = R \begin{pmatrix} 1 & 0 & \langle \widetilde{t}, R(i) \rangle \\ 0 & 1 & \langle \widetilde{t}, R(j) \rangle \\ 0 & 0 & 1 + \langle \widetilde{t}, R(k) \rangle \end{pmatrix}$$

soit $\mathcal{M}_\varphi = RH$.

Réciproquement, on a

$$\begin{aligned} f(x, y) &= (\psi g)(x, y) = g(\psi(x, y)) \\ &= g\left(\frac{ax + a'y + a'' - \langle \widetilde{t}, R(i) \rangle}{cx + c'y + c'' - \langle \widetilde{t}, R(j) \rangle}, \frac{bx + b'y + b'' - \langle \widetilde{t}, R(j) \rangle}{cx + c'y + c'' - \langle \widetilde{t}, R(j) \rangle} \right) \end{aligned}$$

et la matrice $3D$ associée à ψ est

$$\mathcal{M}_\psi = \begin{pmatrix} a & a' & a'' - \langle \widetilde{t}, R(i) \rangle \\ b & b' & b'' - \langle \widetilde{t}, R(j) \rangle \\ c & c' & c'' - \langle \widetilde{t}, R(k) \rangle \end{pmatrix} = R^{-1} \begin{pmatrix} 1 & 0 & -\widetilde{\alpha} \\ 0 & 1 & -\widetilde{\beta} \\ 0 & 0 & 1 - \widetilde{\gamma} \end{pmatrix}.$$

Les expressions de φ et ψ montrent qu'il est vain de chercher à estimer, à partir des images f et g, à la fois la profondeur Z_0 de la scène et la translation de la caméra. En effet, si le vecteur (t, Z_0) est solution des équations précédentes, alors $\lambda(t, Z_0)$ l'est aussi. On ne pourra donc estimer que la direction de la translation.

Les transformations projectives φ et ψ, classiquement représentées dans le groupe projectif à huit paramètres, ne dépendent que des six paramètres du mouvement : trois respectivement pour la rotation et la translation. Il existe un groupe associé aux déplacements et adapté à la modélisation de telles transformations car il est isomorphe au groupe des déplacements rigides de l'espace $SE(3)$; c'est le groupe des recalages, défini dans [1], dans lequel nous nous plaçons. Il est essentiel pour nous de modéliser les déformations d'images dans un groupe car la structure de groupe permet d'inverser et de composer les déformations (et les mouvements de caméra associés).

3 Décomposition du mouvement d'une caméra

Tout d'abord, toute rotation de caméra R peut se décomposer en deux rotations R_1 et R_2 ; la première, R_1, d'axe Δ dans le plan (C, i, j) transforme la direction de l'axe optique k en $R(k)$ tandis que la seconde, R_2, est une rotation autour du nouvel axe $R(k)$. La rotation R_1 dépend de deux paramètres : θ pour la localisation de Δ dans (C, i, j) et α pour l'angle de rotation ; la rotation R_2 est définie par son angle β. En notant R_a^l la rotation d'axe l et d'angle a, on aboutit après calculs à

$$R = R_2 R_1 = R_\theta^k R_\alpha^i R_{-\theta}^k R_\beta^k = R_{\theta,\alpha} R_\beta^k,$$

avec $R_{\theta,\alpha} = R_\theta^k R_\alpha^i R_{-\theta}^k$.

Pour un mouvement de caméra complet, nous pouvons alors écrire

$$RH = R_{\theta,\alpha} R_\beta^k H.$$

Cette décomposition est intéressante par les transformations qu'elle induit sur l'image : la déformation « purement » projective $r_{\theta,\alpha}$ générée par $R_{\theta,\alpha}$ et la similitude s produite par $R_\beta^k H$. On a alors

$$g = \varphi f = s(r_{\theta,\alpha} f).$$

En conséquence, les six paramètres définissant un mouvement de caméra sont répartis comme suit : deux pour la rotation $R_{\theta,\alpha}$ et quatre pour la translation T et la rotation R_β^k. Nous définirons dorénavant un mouvement de caméra par les paramètres $(\theta, \alpha, \beta, A, B, \lambda)$ où $(-A, -B, \lambda - 1)$ sont les coordonnées de $\widetilde{t}$ dans la base $(R(i), R(j), R(k))$. Cette notation allège l'écriture de l'application projective ψ, l'inverse de φ dans le groupe des recalages

$$\psi(x, y) = \left(\frac{ax + a'y + a'' + A}{cx + c'y + c'' + 1 - \lambda}, \frac{bx + b'y + b'' + B}{cx + c'y + c'' + 1 - \lambda} \right).$$

Le tableau (1) donne les ordres de grandeur des valeurs des paramètres, obtenus expérimentalement en prenant des images variées, de dimensions quelques centaines de pixels, auxquelles nous avons appliqué des transformations projectives définies par les six paramètres. Les valeurs retenues permettent de générer des séquences visuellement réalistes, c'est-à-dire sans impression de saccades visuelles entre les images.

TAB. 1 – Ordres de grandeur des valeurs des paramètres entre deux acquisitions d'images consécutives

paramètre	ordre de grandeur
θ (radian)	$]-\pi,\pi]$
α (radian)	10^{-4}
$\|\beta\|$ (radian)	10^{-2}
$\|A\|$,$\|B\|$ (en pixel)	1
$\|1-\lambda\|$ (en pixel)	10^{-2}

En tenant compte des valeurs données dans le tableau, on obtient, en effectuant des développements de Taylor d'ordre 1 en α, d'ordre 2 en β et en $(1-\lambda)$ sur l'expression de ψ, une expression du flot optique à une précision de 10^{-4} sur les valeurs des six paramètres. Puis, en tenant compte de la dimension des images (quelques centaines de pixels), on aboutit à une approximation quadratique du flot optique en un point (x,y) de f, à une précision de 10^{-1} pixels

$$\left\{\begin{array}{lllll} x'-x \simeq & (\lambda-1)x+A & +\beta y & +\alpha x(y\cos\theta - x\sin\theta) \\ y'-y \simeq & \underbrace{(\lambda-1)y+B}_{(a)} & \underbrace{-\beta x}_{(b)} & \underbrace{+\alpha y(y\cos\theta - x\sin\theta)}_{(c)}. \end{array}\right.$$

Ainsi, le flot optique est décomposé en trois termes indépendants : (a) est dû à la translation, (b) à la rotation R_β^k et (c) à la rotation $R_{\theta,\alpha}$. Ce dernier terme n'intervient que lorsque les termes quadratiques (x^2,y^2, xy) sont au moins d'ordre 10^3. Ceci signifie que pour des images de quelques centaines de pixels, la déformation « purement » projective $r_{\theta,\alpha}$ n'est visible qu'en périphérie de l'image. Le centre de l'image, qui correspond aux plus faibles valeurs de x et y (d'ordre 1 et 10), est principalement déformé par la similitude.

4 Algorithme

Pour f et g deux images successives d'une séquence et $(\theta,\alpha,\beta,A,B,\lambda)$ les six paramètres du mouvement entre les deux acquisitions, l'approximation précédente du flot optique donne, pour des valeurs de $|x|$ et $|y|$ restreintes aux ordres 1 et 10

$$\left\{\begin{array}{l} x' \simeq \lambda x + A + \beta y \\ y' \simeq \lambda y + B - \beta x. \end{array}\right.$$

En conséquence, le centre de l'image f est déformé principalement par la similitude. Cette remarque est à la base de notre méthode d'estimation de mouvement de caméra. Les étapes de l'algorithme proposé sont les suivantes :

(1) estimation des paramètres de la similitude $(\hat\beta,\hat A,\hat B,\hat\lambda)$ à partir des zones centrales des images f et g et en utilisant l'algorithme de mouvement paramétrique 2D d'Odobez et Bouthémy [5], disponible à l'adresse
http ://www.irisa.fr/Vista/Motion2D.
Nous utilisons un modèle de mouvement affine à quatre paramètres, permettant de modéliser les translations et rotations 2D et les homothéties. Ceci permet de calculer V_s, le flot optique généré par la similitude définie par les paramètres $(\hat\beta,\hat A,\hat B,\hat\lambda)$.

(2) calcul du flot optique V entre les images f et g par l'algorithme de Weickert et Schnörr [9],

(3) calcul du flot optique $V_p = V - V_s$.
Le flot V_s étant le flot dû à la similitude estimée à l'étape (1), notre approximation permet de considérer que le flot V_p est dû à la transformation « purement » projective (car $V \simeq V_s + V_p$).

(4) estimation du paramètre θ.
La matrice de la rotation $R_{\theta,\alpha}$ est approximée, à une précision de 10^{-4} sur les valeurs des coefficients par

$$R_{\theta,\alpha} \simeq \begin{pmatrix} 1 & 0 & \alpha\sin\theta \\ 0 & 1 & -\alpha\cos\theta \\ -\alpha\sin\theta & \alpha\cos\theta & 1 \end{pmatrix}.$$

Le flot optique généré par la rotation $R_{\theta,\alpha}$ seule, au point (x,y), est ainsi approximé par

$$\left\{\begin{array}{l} x'-x \simeq \dfrac{\alpha(-x^2\sin\theta + xy\cos\theta - \sin\theta)}{\alpha(x\sin\theta - y\cos\theta)+1} \\ \\ y'-y \simeq \dfrac{\alpha(-xy\sin\theta + y^2\cos\theta + \cos\theta)}{\alpha(x\sin\theta - y\cos\theta)+1} \end{array}\right. .$$

On observe alors que la direction $D_p(x,y) = \arctan(\frac{y'-y}{x'-x})$ du vecteur de flot est indépendante de α. On simule donc huit modèles de directions et on compare les directions D_{θ_k} dues aux rotations $R_{\theta_k,\alpha}$ ($\theta_k = -\pi + \frac{k\pi}{4}$, $1 \le k \le 8$ et α arbitrairement choisi) à notre estimation D_p. La valeur sélectionnée minimise un critère L^2

$$\hat\theta = \mathrm{argmin}_{\theta_k} \sum_{(x,y)\in f} (D_p(x,y) - D_{\theta_k}(x,y))^2.$$

(5) estimation de l'angle α.
On construit les images $s(r_{\hat\theta,\alpha_i} f)$, s étant la similitude générée par $(\hat\beta,\hat A,\hat B,\hat\lambda)$ et $r_{\hat\theta,\alpha_i}$ la déformation générée par $R_{\hat\theta,\alpha_i}$, avec α_i appartenant à un ensemble discret et $\hat\theta$ la valeur estimée à l'étape (4). La valeur α_i sélectionnée minimise la différence en norme L^2 $\|g - s(r_{\hat\theta,\alpha_i} f)\|^2$; elle correspond donc au meilleur recalage de f sur g connaissant une estimation des cinq autres paramètres du mouvement.

Dans la méthode présentée, on peut interpréter l'estimation de la déformation purement projective comme une étape de raffinement après l'estimation de la similitude, déformation principale du centre de l'image.

Remarquons que nous recherchons la similitude en utilisant les images et non le flot optique car le flot optique calculé sur une séquence est généralement assez bruité et comme cette première étape conditionne les étapes suivantes, il est préférable de limiter les erreurs d'estimation.

Il est également possible d'estimer directement les paramètres du mouvement de la caméra à partir des deux images en utilisant une autre version de l'algorithme d'estimation de mouvement paramétrique d'Odobez et Bouthémy. Le modèle 2D utilisé est alors un modèle quadratique à huit paramètres, que l'on sait regrouper pour obtenir

les six paramètres du mouvement de la caméra, grâce à la décomposition précédente du flot optique en les trois termes (a), (b) et (c).

5 Résultats et conclusion

Sur une séquence simulée à partir d'une image, nous obtenons de bons résultats ; le mouvement de la caméra est bien estimé. De plus, on a observé que l'étape d'estimation de l'angle θ est particulièrement robuste à un bruit impulsionnel ajouté sur les composantes du flot optique.

Nous avons également appliqué la méthode présentée à un film de notre bureau. En combinant les petits mouvements estimés entre les images successives, on accède au mouvement de la caméra entre des images éloignées dans le temps. Ceci nous permet de réaliser des panoramas, car la scène est suffisamment éloignée du centre optique. De tels mosaïquages sont donnés sur les figures (1) et (2). Remarquons que des discontinuités apparaissent lorsque les zones recalées ne correspondent pas à la profondeur moyenne de la scène. Notre prochain objectif est d'utiliser ces informations et le mouvement de la caméra pour déterminer une carte des disparités de la scène.

FIG. 1 – De gauche à droite, les images 70, 80 et 97 de la séquence du bureau et l'image reconstituée du point de vue de l'image 80.

FIG. 2 – De gauche à droite, les images 20, 35 et 50 de la séquence du bureau et l'image reconstituée du point de vue de l'image 35.

Références

[1] F. Dibos, G. Koepfler, P. Monasse, *Image Alignment*, Geometric Level Set Methods in Imaging, Vision and Graphics, Springer, 2003.

[2] O. Faugeras, *Three-dimensional computer vision : a geometric viewpoint*, MIT Press, 1993.

[3] D.J. Heeger, A.D. Jepson, *Subspace methods for recovering rigid motion i : algorithm and implementation*, IJCV, 7(2) :95-117, 1992.

[4] J.E. Ha, I.S. Kweon, *Robust direct motion estimation considering discontinuity*, Pattern Recognition Letters, 21(11) :999-1011, 2000.

[5] J.M. Odobez, P. Bouthémy, *Robust Multiresolution Estimation of Parametric Motion Models*, Jal. of Visual Communication and Image Representation, 6(4) :348-365, 1995.

[6] Y. Ma, J. Koseckà, S. Sastry, *Linear differential algorithm for motion recovery : A geometric approach*, IJCV, 36(1) :71-89, 2000.

[7] J.H. Rieger, D.T. Lawton, *Processing differential image motion*, Journal of the Optical Society of America, 2(2) :354-360, 1997.

[8] Y. Tian, C. Tomasi, D.J. Heeger, *Comparison of approaches to egomotion computation*, IEEE, Conference on Computer Vision and Pattern Recognition, 315-320, 1996.

[9] J. Weickert, C. Schnörr, *Variational Optic Flow Computation with a Spatio-Temporal Smoothness Constraint*, Technical Report, Computer Science Series, 2000.