Extraction d'un modèle 3D de visage en temps-réel et de manière robuste

Marc CHAUMONT

LIRMM, UMR CNRS 5506, Université de Montpellier II,
161 rue ADA, 34392 Montpellier Cedex 05, France.
marc.chaumont@lirmm.fr

Résumé – Cet article traite de l'extraction d'un modèle 3D de visage et de son positionnement 3D à partir d'un faible nombre de couples de points 2D-3D en correspondance. Cette extraction est envisagée comme une phase d'initialisation pour une application de suivi de visage temps-réel basé modèle 3D. L'inconvénient majeur des solutions actuelles pour l'extraction de modèle 3D sont soit la complexité calculatoire, soit la modélisation simplifiée. Notre solution, quant à elle, est rapide, robuste et suffisamment descriptive, ce qui la rend exploitable comme initialisation pour un suivi de visage temps-réel basé modèle 3D. Cette solution se décompose en deux étapes : une approximation suivie d'un raffinement d'un modèle 3D générique. Les résultats obtenus montre des performances en terme de robustesse, de rapidité et de réalisme.

Abstract – This article deals with 3D-face model and 3D-pose extraction from a small set of couples of 2D-3D corresponding-points. This extraction is considered as an initialization step of a real-time face tracking application based on a 3D model. Major drawbacks of current 3D model extraction solutions are either the computationally complexity or the over-simplified modeling. Our solution is rapid, robust and descriptive enough, which make it useful for a face tracking, in real-time, based on a 3D model. This solution is based on a two step approach: an approximation followed by a refinement of a generic 3D model. The results obtained show rapid, robust and realistic performances.

1 Introduction

Il existe deux grandes classes de techniques d'extraction de modèle 3D : la première classe nécessite un minimum de deux images pour faire un calcul de triangulation et retrouver les positions 3D des points utilisés lors de la triangulation [1] ; la seconde classe nécessite la connaissance d'un modèle 3D qu'il faut positionner [2]. Notons qu'une contrainte supplémentaire s'ajoute lors de l'extraction d'un modèle de visage, puisque les modèles 3D sont génériques et doivent être déformés pour être adaptés à la morphologie et aux émotions d'un visage particulier.

Ces deux approches sont coûteuses en temps de calcul et nécessitent pour être robustes de disposer d'un grand nombre de points 2D. Or, dans le cadre du suivi de visage par modèle 3D, lorsque les séquences vidéo sont acquises par une caméra non calibrée, la phase d'initialisation (c'est-à-dire d'extraction et de positionnement du modèle 3D) doit être à la fois robuste et temps-réel. La solution que nous proposons permet d'extraire en temps-réel un modèle 3D à partir d'un faible nombre de points 2D caractérisant le visage.

Le schéma général que nous proposons se base sur l'utilisation d'un modèle 3D générique et se compose de deux étapes. La première étape détaillée en partie 3 consiste à extraire une approximation du modèle 3D de visage. La seconde étape détaillée en partie 4 vise à améliorer le modèle 3D et à extraire sa position 3D. Pour introduire les deux parties relatives à notre contribution, nous formalisons au préalable le problème dans la partie 2.

2 Formulation énergétique générale

Avec une caméra classique de type sténopé, la projection T d'un point 3D $M'_i = (X'_i, Y'_i, Z'_i)^t$ (dans le repère objet) donne un point 2D $m'_i = (u'_i, v'_i)^t$ (dans le repère image) qui peut être exprimé en coordonnées homogènes [3] par l'équation 1. f est la longueur focale de la caméra ; k_u et k_v sont les facteurs d'échelle horizontaux et verticaux (mesurés en pixels/m) ; u_0 et v_0 sont les coordonnées du point principal ; $(t_x, t_y, t_z)^t$ est le vecteur de translation et $r_{ij:i,j\in[1,3]}$ sont les coeffi cients de la matrice de rotation. La Figure 1 illustre les différents systèmes de coordonnées ainsi que la projection d'un point 3D M'_i en un point 2D m'_i.

Formellement, l'extraction et le positionnement du modèle 3D de visage consiste à minimiser la distance E (voir l'équation 2) entre l'ensemble observé des points 2D de l'image $\{(u_i, v_i)^t\}$ et l'ensemble des points projetés $\{(u'_i, v'_i)^t\}$. L'ensemble des points projetés $\{(u'_i, v'_i)^t\}$ est obtenu en projetant les sommets du modèle 3D de visage, correspondant aux points 2D observés, en utilisant la projection T de l'équation 1.

$$E = \sum_i (u_i - u'_i)^2 + (v_i - v'_i)^2. \qquad (2)$$

Notons que les sommets du modèle 3D de visage utilisés dans la minimisation de l'équation 2 appartiennent à un modèle 3D de visage "*mis-en-forme*". Par "*mis-en-forme*", nous voulons dire que la morphologie et les émotions courantes du visage traité sont décrites par le modèle. Pour obtenir ce modèle 3D "*mis-en-forme*" nous déplaçons les sommets d'un modèle 3D générique. Le modèle 3D que nous avons retenu se nomme CANDIDE-3 [4] ; il décrit un visage moyen en position neutre par un maillage composé de 113 sommets et 168 facettes.

L'ensemble des points projetés $\{(u'_i, v'_i)^t\}$ (voir l'équation

$$\begin{pmatrix} s.u'_i \\ s.v'_i \\ s \end{pmatrix} = \underbrace{\begin{pmatrix} \alpha_u & 0 & u_0 & 0 \\ 0 & \alpha_v & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}}_{paramètres\ intrinsèques} . \underbrace{\begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{paramètres\ extrinsèques} . \begin{pmatrix} X'_i \\ Y'_i \\ Z'_i \\ 1 \end{pmatrix} = T_{3\times4}.M'_i,\ avec \begin{cases} \alpha_u = -k_u.f \\ \alpha_v = k_v.f \end{cases} . \quad (1)$$

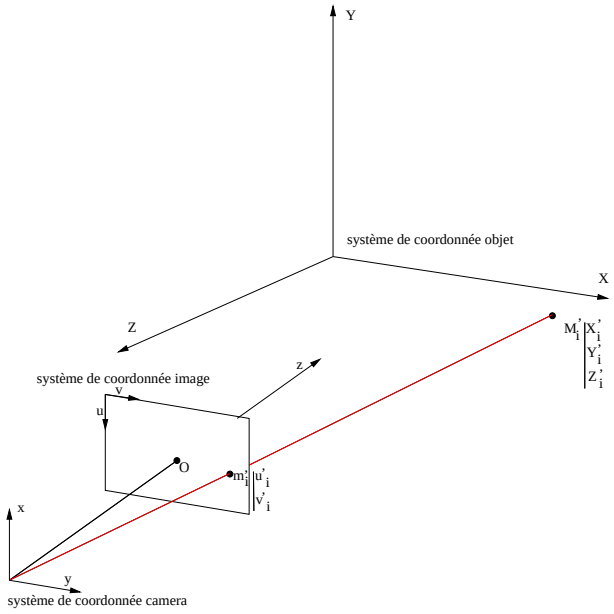


FIG. 1 – Les différents systèmes de coordonnée

2) est obtenu par l'application de trois opérations consécutives sur les sommets du modèle CANDIDE-3. Ces trois opérations

$$\begin{pmatrix} s.u'_i \\ s.v'_i \\ s \end{pmatrix} = T.[\underbrace{M_i + S_i\sigma + A_i.\alpha}_{M'_i}]. \quad (3)$$

sont : **1.** un *déplacement de forme* $S_i.\sigma$; **2.** un *déplacement d'animation* $A_i.\alpha$; **3.** la projection T. Ceci est exprimé par l'équation 3. S_i et A_i sont respectivement la *matrice d'unité de forme* et la *matrice d'unité d'animation*, exprimant le déplacement possible d'un sommet i du modèle CANDIDE-3. L'intensité du déplacement est exprimée par les vecteurs de pondération σ et α. Pour plus de détails, le lecteur pourra consulter le rapport d'Ahlberg [4]. La minimisation de l'équation 2 permet d'obtenir les paramètres T, σ et α.

Si elle est traitée directement, la minimisation de l'équation 2 est diffi cile, la solution n'est pas obtenue en temps-réel, et le résultat est peu robuste. Pour parvenir à une solution acceptable en peu de temps, nous proposons une solution qui se décompose en deux étapes. La première étape (partie 3) consiste à approximer la forme du modèle 3D de visage (T, σ et α sont grossièrement calculés). La seconde étape (partie 4) permet d'améliorer la forme du modèle 3D et d'extraire sa position 3D (T et la forme du modèle 3D sont raffi nés).

3 Approximation du modèle 3D de visage

3.1 Approximation de la pose

Le calcul de la projection T (obtenue en résolvant l'équation 2) n'est pas une tâche facile. L'expression immédiate de la solution mène en effet à un système linéaire homogène. Pour résoudre l'équation, il est classique d'introduire des contraintes sur la solution comme dans [5]. Pour cela, la connaissance d'un grand nombre de couples de points 2D-3D en correspondances est nécessaire. Or, dans le cas d'un visage Humain, le nombre de points caractéristiques est faible. Le nombre de couples de points 2D-3D en correspondance n'est donc pas suffi sant pour utiliser ce type d'approche. Nous proposons donc de simplifi er la projection T.

Cette simplifi cation consiste à supposer que tous les sommets 3D sont dans un même plan. Cette hypothèse est réaliste lorsque la distance entre la caméra et le visage est grande en comparaison des différences de profondeurs entre les points 3D du visage. La projection T est donc simplifi ée en une projection perspective faible[1] dont la matrice associée est de dimension 2×4 tel que :

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} (\alpha_u.r_{1j} + u_0.r_{3j})/t_z & \alpha_u.t_x/t_z + u_0 \\ (\alpha_v.r_{2j} + v_0.r_{3j})/t_z & \alpha_v.t_y/t_z + v_0 \end{pmatrix}_{j\in[1,3]} . \begin{pmatrix} M_i \\ 1 \end{pmatrix}$$
$$= \underbrace{\begin{pmatrix} a_0 & b_0 & c_0 & d_0 \\ a_1 & b_1 & c_1 & d_1 \end{pmatrix}}_{T_{2\times4}} . \begin{pmatrix} M_i \\ 1 \end{pmatrix} .$$

En annulant, à partir de l'équation 2, chacune des dérivées partielles de E fonction des paramètres de T, nous obtenons deux systèmes linéaires (σ et α sont positionnés à zéro). Le premier système linéaire est donné Équation 4 (le deuxième système linéaire est obtenu en remplaçant u_i par v_i et a_0 par a_1, b_0 par b_1 et ainsi de suite). Les deux systèmes sont résolus en utilisant des outils classiques d'algèbre linéaire.

$$\begin{pmatrix} \sum_i X_i^2 & \sum_i X_i.Y_i & \sum_i X_i.Z_i & \sum_i X_i \\ \sum_i X_i.Y_i & \sum_i Y_i^2 & \sum_i Y_i.Z_i & \sum_i Y_i \\ \sum_i X_i.Z_i & \sum_i Y_i.Z_i & \sum_i Z_i^2 & \sum_i Z_i \\ \sum_i X_i & \sum_i Y_i & \sum_i Z_i & \sum_i 1 \end{pmatrix} . \begin{pmatrix} a_0 \\ b_0 \\ c_0 \\ d_0 \end{pmatrix}$$
$$= \begin{pmatrix} \sum_i X_i.u_i \\ \sum_i Y_i.u_i \\ \sum_i Z_i.u_i \\ \sum_i u_i \end{pmatrix} . \quad (4)$$

3.2 Approximation de la forme

Une fois que la projection $T_{2\times4}$ est calculée, nous pouvons extraire le modèle 3D du visage. Pour cela, nous allons "*mettre-en-forme*" le modèle 3D moyen *CANDIDE-3*. La minimisation de l'équation 2 est effectuée en fi xant la projection $T_{2\times4}$. L'équation 2 est réécrite telle que :

[1] La projection perspective faible est également appelée projection orthographique, projection orthographique graduée ou projection affi ne.

$$E = \sum_i [U_i - N.S_i.\sigma]^t.[U_i - N.S_i.\sigma], \quad (5)$$
$$avec\, U_i = \begin{pmatrix} u'_i \\ v'_i \end{pmatrix} - T. \begin{pmatrix} M_i \\ 1 \end{pmatrix},$$
$$et \quad N = T_{2\times 3}.$$

On obtient alors un système linéaire (équation 6) en annulant les dérivées partielles $\frac{\partial E}{\partial \sigma}$. La solution est telle que $\sigma = (A^t A)^{-1} A^t.B$.

$$\underbrace{(\sum_i S_i^t.N^t.N.S_i)}_{A}.\sigma = \underbrace{\sum_i S_i^t.N^t.U_i}_{B}. \quad (6)$$

Du fait du faible nombre de couples de points 2D-3D en correspondance, la matrice A est souvent non inversible. On la re-conditionne donc de manière à ce que les coeffi cients diagonaux à l'intersection d'une ligne de zéros et d'une colonne de zéros soient non nuls. Notons également que la solution du système linéaire ne doit pas mener à des retournements de maille 3D ni à des déplacements de grande amplitude par rapport à la taille du modèle 3D. Pour cela, on force les coeffi cients de σ à appartenir à leur intervalle de défi nition $[-1, 1]$. Enfi n, pour plus de robustesse, seuls les points peu animés ou non animés sont considérés pour la construction de la matrice A et du vecteur B. Le même raisonnement peut être tenu pour effectuer le calcul de α.

4 Raffinement du modèle 3D et de la pose 3D

Dans la partie précédente, nous avons expliqué comment obtenir rapidement une pose 3D approximative (la matrice $T_{2\times 4}$) et un modèle 3D approximatif. Notre objectif est maintenant d'extraire une information de position 3D plus descriptive (en particulier l'information de profondeur : t_z) et d'avoir un modèle 3D plus précis (c'est-à-dire avoir une meilleure mise en correspondance entre points 2D et points 3D projetés).

4.1 Extraction de la position 3D et raffinement du modèle 3D

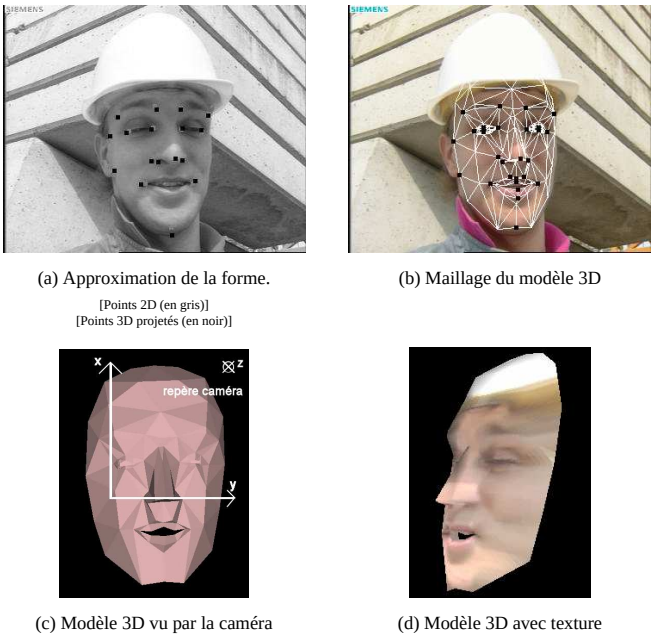
Pour extraire les paramètres extrinsèques (la rotation et la translation), nous utilisons l'algorithme *POSIT*[2] de DeMenthon [6] en fi xant les paramètres intrinsèques[3]. Ce choix est guidé par le fait que les paramètres intrinsèques peuvent être approximés sans grande erreur de reconstruction [7, 8] et parceque l'algorithme *POSIT* est simple et rapide.

Des erreurs de mise en correspondance sont introduites : **1.** lors de l'approximation de T (sous-partie 3.1), **2.** lors de la déformation du modèle 3D par l'utilisation d'unités de forme et d'animation (sous-partie 3.2), **3.** lors du positionnement des paramètres intrinsèques. Pour obtenir une mise en correspondance précise nous déplaçons séparément chaque sommet 3D du modèle et pour plus de robustesse, nous déplaçons uniquement les sommets 3D dont l'erreur de mise en correspondance est faible. Le déplacement de chaque sommet M_i est obtenu en résolvant l'équation linéaire 7 (X_i et Y_i sont les inconnus) pour les couples de points valides. Remarquons que cette dernière étape doit être mené uniquement si l'erreur moyenne de mise en correspondance est faible.

$$\begin{pmatrix} ((u_i - u_0).r_{31} - \alpha_u.r_{11}) & ((u_i - u_0).r_{32} - \alpha_u.r_{12}) \\ ((v_i - v_0).r_{31} - \alpha_v.r_{21}) & ((v_i - v_0).r_{32} - \alpha_v.r_{22}) \end{pmatrix} \cdot \begin{pmatrix} X_i \\ Y_i \end{pmatrix}$$
$$= \begin{pmatrix} (\alpha_u.r_{13} - (u_i - u_0).r_{33}).Z_i + \alpha_u.t_x - (u_i - u_0).t_z \\ (\alpha_v.r_{23} - (v_i - v_0).r_{33}).Z_i + \alpha_v.t_y - (v_i - v_0).t_z \end{pmatrix}. \quad (7)$$

5 Résultats

La Figure 2 illustre sur l'image *Foreman* quelques étapes de notre technique d'extraction et de positionnement d'un modèle 3D. La fi gure 2(a) donne le résultat de la "*mise-en-forme*" du modèle 3D moyen *CANDIDE-3* ($T_{2\times 4}$ et σ sont calculés). La fi gure 2(b) montre le maillage fi nal du modèle 3D de visage projeté. On peut constater que le modèle représente très bien la morphologie et l'animation du visage. Les fi gures 2(c) et 2(d) montrent que l'utilisation d'un modèle 3D pour le suivi de visage est envisageable puisqu'il est possible de synthétiser de manière réaliste le visage sous un angle de vue différent, ceci avec un faible nombre de points en entrée, un modèle 3D possédant peu de facettes, et un temps de calcul très faible.



(a) Approximation de la forme. [Points 2D (en gris)] [Points 3D projetés (en noir)]

(b) Maillage du modèle 3D

(c) Modèle 3D vu par la caméra

(d) Modèle 3D avec texture

FIG. 2 – Illustration de certaines étapes de l'extraction et du positionnement du modèle 3D sur l'image *Foreman*

Le tableau 1 donne les erreurs moyennes de mise en correspondance entre les points 2D placés manuellement et les points 3D projetés. L'erreur est donnée pour le modèle projectif perspective faible : $T_{2\times 4}$ et pour le modèle projectif sténopé : $T_{3\times 4}$ (paramètres intrinsèques fi xés ; paramètres extrinsèques calculés par l'algorithme *POSIT*). On peut constater qu'en règle gé-

[2] L'algorithme *POSIT* consiste à réduire de manière itérative l'erreur de projection due à l'approximation de la projection de l'équation 1 par une projection perspective faible, ceci dans l'objectif d'extraire les paramètres extrinsèques.

[3] f est positionnée à 0.05, k_u et k_v sont positionnés à 5000, le point $(u_0, v_0)^t$ est positionné au centre de l'image

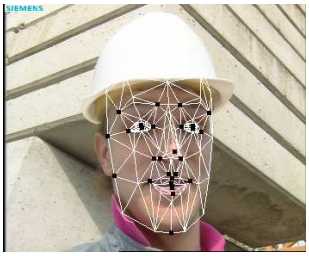
nérale le modèle projectif faible est plus précis que le modèle sténopé.

TAB. 1 – Erreurs moyennes de mise en correspondance avec la projection perspective faible et la projection sténopée

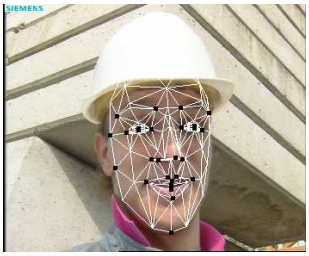
Image	Nb. points	Type de projection	Erreur
Foreman	27	$T_{2\times4}$	3.09
		$T_{3\times4}$	2.88
Lena	23	$T_{2\times4}$	2.78
		$T_{3\times4}$	3.15
Rotation (base M2VTS)	15	$T_{2\times4}$	1.62
		$T_{3\times4}$	5.22

La figure 3 illustre le cas où l'erreur de mise en correspondance est réduite lors du passage au modèle sténopé T_{3x4}. La figure 4 illustre le cas inverse avec une erreur de mise en correspondance fortement augmentée lors du passage au modèle sténopé.

L'erreur obtenue par le modèle sténopé peut s'expliquer par l'imprécision sur les paramètres intrinsèques et extrinsèques. Il pourrait donc être envisagé d'ajouter à l'approche le calcul des paramètres intrinsèques : soit en utilisant l'auto-calibration basée analyse-synthèse [9], soit en utilisant des approches provenant du domaine de la vision [10]. Le calcul des paramètres intrinsèques par de telles approches reste cependant sensible à la phase d'initialisation ainsi qu'aux bruits sur les points 2D. Le calcul des paramètres extrinsèque peut également être amélioré. L'algorithme *POSIT* fournit en effet une approximation des paramètres extrinsèques et une approche linéaire robuste [11] pourrait s'y substituer.



(c) Projection $T_{2\times4}$



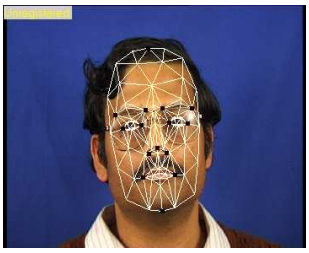
(d) Projection $T_{3\times4}$

FIG. 3 – Illustration de la projection du maillage 3D, sur l'image *Foreman*, avec un modèle projectif à perspective faible $T_{2\times4}$ et avec un modèle projectif sténopé $T_{3\times4}$

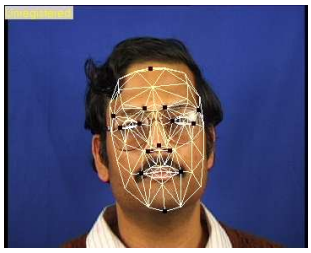
Le temps d'exécution nécessaire à l'extraction du modèle 3D et de sa pose, à partir d'une trentaine de points 2D, sur un processeur Intel Pentium cadencé à 2.4Ghz est de 1.6 ms (l'implémentation est en C++ et lors de la mesure, une dizaine d'autres applications fonctionnaient). Ce temps de calcul peut par ailleurs être réduit par optimisation du code et utilisation d'une librairie de calcul matriciel plus performante.

6 Conclusion

Dans ce papier nous proposons une solution robuste et rapide pour positionner et extraire un modèle de 3D de visage. Dans ce but, nous utilisons un modèle 3D moyen de visage et quelques couples de points 2D-3D en correspondance. Une succession de calculs robustes et rapides mènent à un modèle 3D reflétant bien la structure du visage traité, et à un positionnement du modèle 3D très complet puisque l'on détermine les paramètres du modèle projectif de caméra sténopé (paramètres de caméra, de rotation et de translation).



(a) Projection $T_{2\times4}$



(b) Projection $T_{3\times4}$

FIG. 4 – Illustration de la projection du maillage 3D, sur une image d'une séquence de la base M2VTS, avec un modèle projectif à perspective faible $T_{2\times4}$ et avec un modèle projectif sténopé $T_{3\times4}$

Références

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

[2] F. Dornaika and J. Ahlberg, "Fast and reliable active appearance model search for 3D face tracking," *Transactions on Systems, Man, and Cybernetics–Part B : Cybernetics*, vol. 34, no. 4, pp. 1838–1853, 2004.

[3] R. Horaud et O. Monga, *Vision par ordinateur : Outils fondamentaux*, chapter 5, Editions Hermès, Eyrolles, 1995.

[4] J. Ahlberg, "CANDIDE-3 - un updated parameterised face," Tech. Rep., Department of Electrical Engineering, Linköping University, Jan. 2001.

[5] O. D. Faugeras and G. Toscani, "The calibration problem for stereo," in *Computer Vision and Pattern Recognition, CVPR*, 1986 June, pp. 15–20.

[6] D. DeMenthon and L.S. Davis, "Model-based object pose in 25 lines of code," *International Journal of Computer Vision, IJCV*, vol. 15, no. 1, pp. 123–141, June 1995.

[7] S. Bougnoux, "From projective to euclidean space under any practical situation, a criticism of self-calibration," in *International Conference on Computer Vision, ICCV*, Jan. 1998, pp. 790–798.

[8] L.F. Cheong and C-H. Peh, "Characterizing depth distortion due to calibration uncertainty," in *European Conference on Computer Vision, ECCV*. June 2000, vol. 1842, pp. 664–677, Springer-Verlag.

[9] P. Eisert, "Model-based camera calibration using analysis by synthesis techniques," in *Proc. 7th International Workshop Vision, Modeling, and Visualization*, Nov. 2002, pp. 307–314.

[10] A. Fusiello, "Uncalibrated euclidean reconstruction : A review," *Image and Vision Computing*, vol. 18, no. 6-7, pp. 555–563, May 2000.

[11] L Quan and Z Lan, "Linear n-point camera pose determination," *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, vol. 21, no. 8, pp. 774–780, Aug. 1999.