Estimation de la pose 3D de la main à partir de séquences d'images monoculaires

Martin de La Gorce,Nikos Paragios,David Fleet

Laboratoire de MAS,Ecole Centrale de Paris
Univeristée de Toronto

15 juin 2007

Objectif

- Entrées :
  - Modèle déformable 3D paramétré de la main
  - Séquence d'images vidéo monoculaires
  - Pose initiale approximative de la main
- Sorties :
  - Paramètres estimés définissant la pose de la main à chaque image
  - Eventuellement : Texture estimée de la main, Orientation lumière

Résultats présentés en 2006

- Modélisation / hypothèses :
  - Volume définit comme union d'ellipsoïdes et de polyèdres rigides
  - Distribution des couleurs de la main modélisée par un seul histogramme
  - Couleur du fond : un processus gaussien différent pour chaque pixel appris en ligne.
- Estimation de la pose de la main par Maximum de vraisemblance
  - les pixels à l'intérieur de la silhouette synthétisée doivent ressembler à l'histogramme de la main
  - les pixels à l'extérieur ressemblent doivent ressemble au fond

Limitations

Les limitations sont surtout dues au model de couleur de la main : La distribution de probabilité de couleur est supposé identique pour tous les pixels dans la main :

- Texture et ombrages non modélisé et donc non pris en compte dans la vraisemblance
- Les bords des auto-occlusions non pris en compte car même modèle de couleur identique de chaque coté
- Conséquence : La fonction coût résultante dépend seulement de la silhouette

Il est possible d'ajouter un terme de distance entre contours de l'image est contours synthétique mais :

- Difficle à justifier dans un cadre Maximum de Vraisemblance
- Discontinuités possibles de la fonction lors d'apparition/disparition d'auto-occlusion

Informations utile dans l'image

- Silhouette et bords peu informative sur la profondeur quand :
  - Les dimensions de l'objet sont inconnues
  - La projection est quasi-orthographique : la taille apparente est constante
- Les ombrages :
  - Contient une information relative à la normale de la surface et donc à la variation de profondeur.
- Texture :
  - Information de déplacement fronto-parallèle dense
  - Seule sopurce d'information dans le cas de rotation autour d'un axe de révolution de l'objet

Toutes ces informations semblent importantes pour le suivit de la main. Il faut donc les modéliser dans notre modèle.

Combiner les sources d'information

Deux approches :

1. Sommation de termes hétérogènes : flot optique, distance de chanfrein etc
   - S. Lu, D. Metaxas, D. Samaras, and J. Oliensis. CVPR 2003 :

   - comment pondérer les différents termes/forces ?
   - Effet d'ombrage et occlusion gênent l'estimation du flot optique
2. Minimiser apparence synthétique et apparence observée
   - "Morpheable model" pour les visages (V. Blanz and T. Vetter, Siggraph 1999)
   - Active Appearance Models

Approche AAM / morphable model

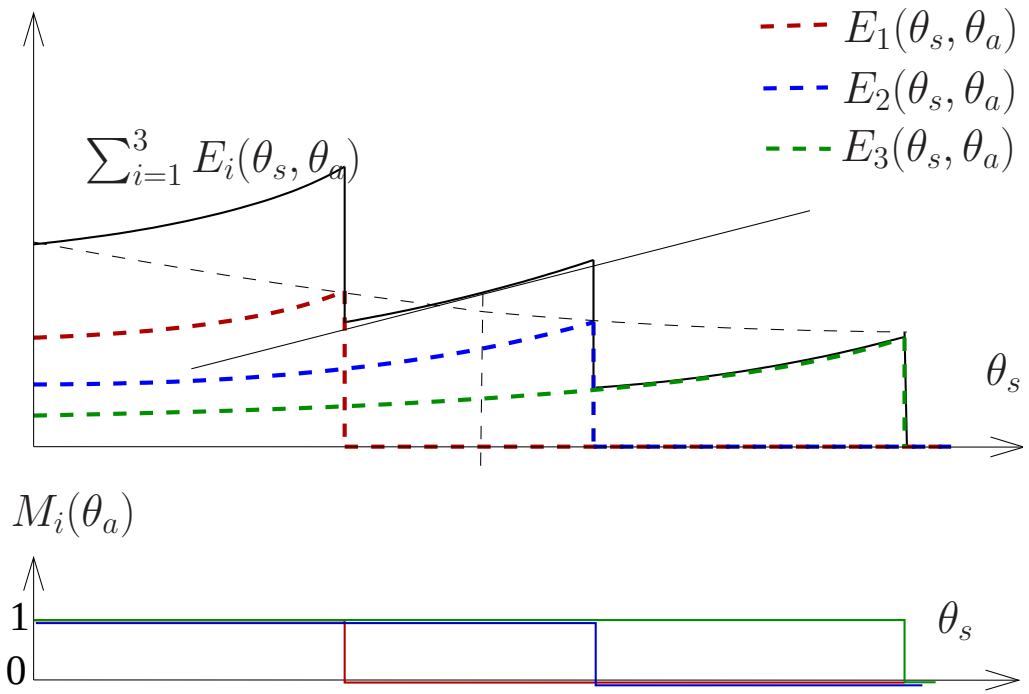
- Forme :
  - définie par un mesh avec N points dans $\mathbb{R}^3$ paramétrés par $\theta_s \in \mathbb{R}_s^N$:
    $$(V_1(\theta_s), \ldots, V_N(\theta_s)) \in (\mathbb{R}^3)^N$$
  - ex : $\theta_s =$ 6 paramètres transformation globale + coefs des modes pour une modèle PCA
- Apparence :
  - $L_i(\theta_s, l) =$ Illumination du point i , dépend de la lumière (l) et de la forme, donc de θ_s
  - $C_i(\theta_a) =$ albedo RGB du point i paramétré par un vecteur $\theta_a \in \mathbb{R}_a^N$
  - ex : $\theta_a =$ coefficient des modes de variation pour une modèle PCA
- Projection : π_{camera}
- Visibilité (surface triangulée) : $M_i(\theta_s)$ variable binaire qui dépend de la pose
- Mesure d'erreur suivant les paramètres de forme θ_s et apparence θ_a :

$$E(\theta_s, \theta_a) = \sum_{i=1}^{N} M_i(\theta_s) \| I_{obs}(\pi_{camera}(V_i(\theta_s))) - L_i(\theta_s) C_i(\theta_a) \|^2$$

Critiques de l'Approche AAM : discontinuité de la visibilité

- Mesure d'erreur discontinue car $M_i(\theta_s)$ binaire
- Visiblité $M_i(\theta_s)$ localement constante : derivée $\frac{\partial M_i(\theta_s)}{\theta_s} = 0$
- Minimaux locaux artificiels
- Direction de descente peu intéressante

$$E_i(\theta_s, \theta_a) = M_i(\theta_a)|I_{obs}(\pi_{camera}(V_i(\theta_s))) - C_i(\theta_a)|^2$$



Solutions envisagées

1. Utiliser une transition continue de la visibilité :
   - $\frac{\partial M_i(\theta_s)}{\theta_s} \neq 0$ donc nouveaux termes dans le gradient
   - Pb : difficile à formaliser la transition
2. Ecrire l'erreur avec une intégrale sur la surface plutôt qu'une somme
   - frontière du masque de visibilité se déplace continument
   - erreur varie continument
   - gradient difficile à calculer
3. Sommer l'erreur en intégrant dans le domaine de l'image :
   - Plus de masque de visibilité
   - Véritable approche **Analyse par synthèse**
   - calcul du gradient relativement élégant ...

On a opté pour la dernière.

Méthode : Analyse par synthèse

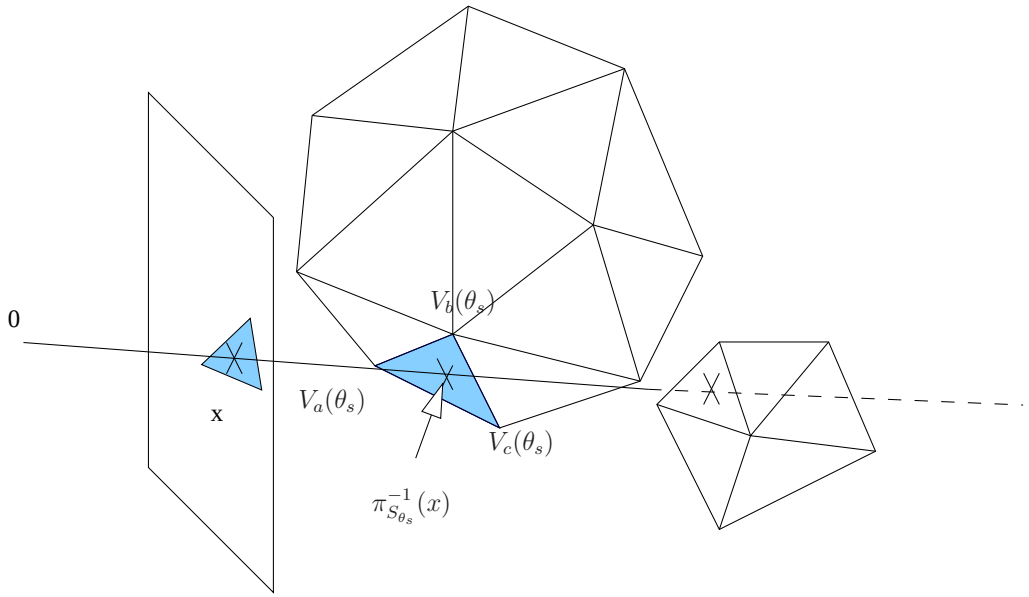
- Synthèse ou création d'une image synthétique $I_{synth}(\theta_s, \theta_a, l)$:
  - Déformation de la surface triangulé suivant les paramètres de déformation θ_s
  - Calcul de l'ombrage pour chaque sommet du mesh suivant θ_s et l
  - Calcul de la texture/apparence à partir des coefficients θ_a
  - Projection des faces texturées dans l'image avec ombrage, texture et auto-occlusions.
- Analyse :
  - Mesure de la différence entre image observé I_{obs} et image synthétisée I_{synth} :

$$E_{dt}(\theta_s, \theta_a, l) = \int_{\Omega} \rho\Big(I_{synth}(\theta_s, \theta_a, l, \mathbf{x}) - I_{obs}(\mathbf{x})\Big) d\mathbf{x} \qquad (1)$$

  - Estimation des paramètres (déformation surface (θ_s) + lumière(l)+ texture(θ_a)) par minimisation de l'erreur avec méthode quasi-newton **utilisant le gradient**

Synthèse : shéma

Formalisée comme un rendu classique d'image de synthèse :



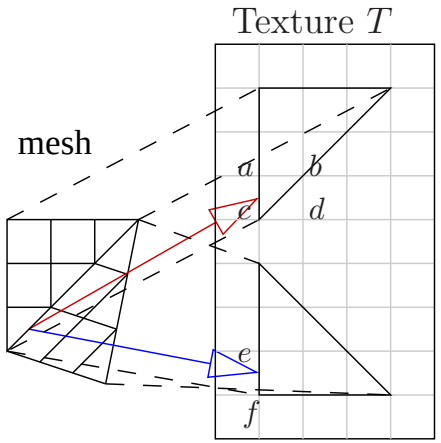
$$*\pi^{-}_{\theta_s} 1(x) = \sum_{i=a,b,c} w_i V_i(\theta_s)) \quad (2)$$

$$I_{synthy}(x) = (\sum_{i=a,b,c} w_i L_i(\theta_s, I))(\sum_{i=a,b,c} w_i C_i(\theta_a)) \quad (3)$$

En pratique pour l'implémentation, pas de lancé de rayon, mais technique de rastérisation avec z-buffer.

Synthèse : texture

- Plutot que d'avoir une couleur RGB par sommet on peux définir une texture avec un mapping de la surface vers la texture et réduire le nombre de faces.
- On souhaite que $I_{synth}(\theta_s, \theta_a, I, \mathbf{x})$ soit une fonction continue par rapport à θ_s si l'on est pas sur le contour d'une (auto-)occlusion :

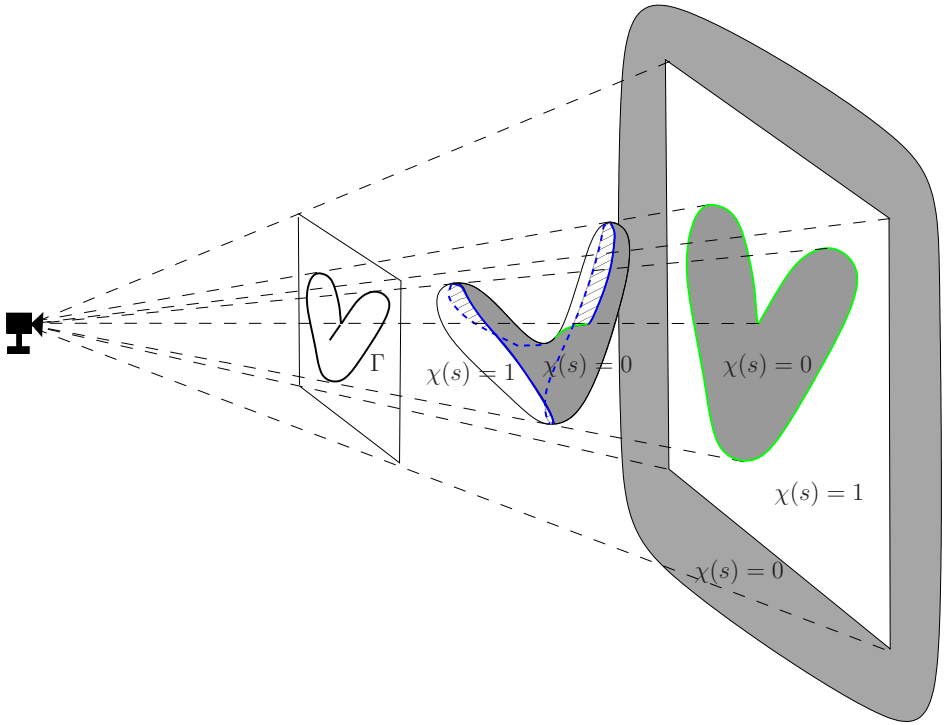


  - Interpolation bilinéaire de la texture
  - Contraintes sur la texture pour éviter les discontinuités entre faces adjacentes :
    $T(a) = T(e)$, $T(c) = T(f)$,
    $\frac{1}{4}(T(a) + T(b) + T(c) + T(d)) = \frac{1}{4}(T(e) + T(f))$

Analyse : contour des occlusions

Pour formaliser le calcul du gradient $\frac{\partial E_{dt}}{\partial \theta_s}$ on définit :

- $\Gamma_{\theta_s} \subset \mathbb{R}^2$ l'ensemble des courbes délimitant les (auto-)occlusions :



La surface étant triangulée :
Γ_{θ_s} est la projection le l'ensemble des points visible appartenant à un bord adjacent à deux faces dont les normales sont orientée de façon opposées relativement à la caméra.

Analyse : déplacement du contour

soit $x \in \Gamma_{\theta_s}$.

- x correspond à la projection d'un point y visible sur un bord entre deux sommet $V_a(\theta_s)$ et $V_b(\theta_s)$
  $x = \pi_{camera}(y),\ y = (1-t)V_a(\theta_s) + tV_b(\theta_s)$
- Le déplacement infinitésimal (avec composante tangentielle) du contour Γ_{θ_s} autour de x quand θ_s varie est décrit par une matrice $v(x, \theta_s)$ de taille 2 par N_s(#para de déformation)
- $v(x, \theta_s) \times \Delta\theta_s \sim$ déplacement de la courbe autour de x lorsque l'on passe de θ_s à $\theta_s + \Delta\theta_s$
-

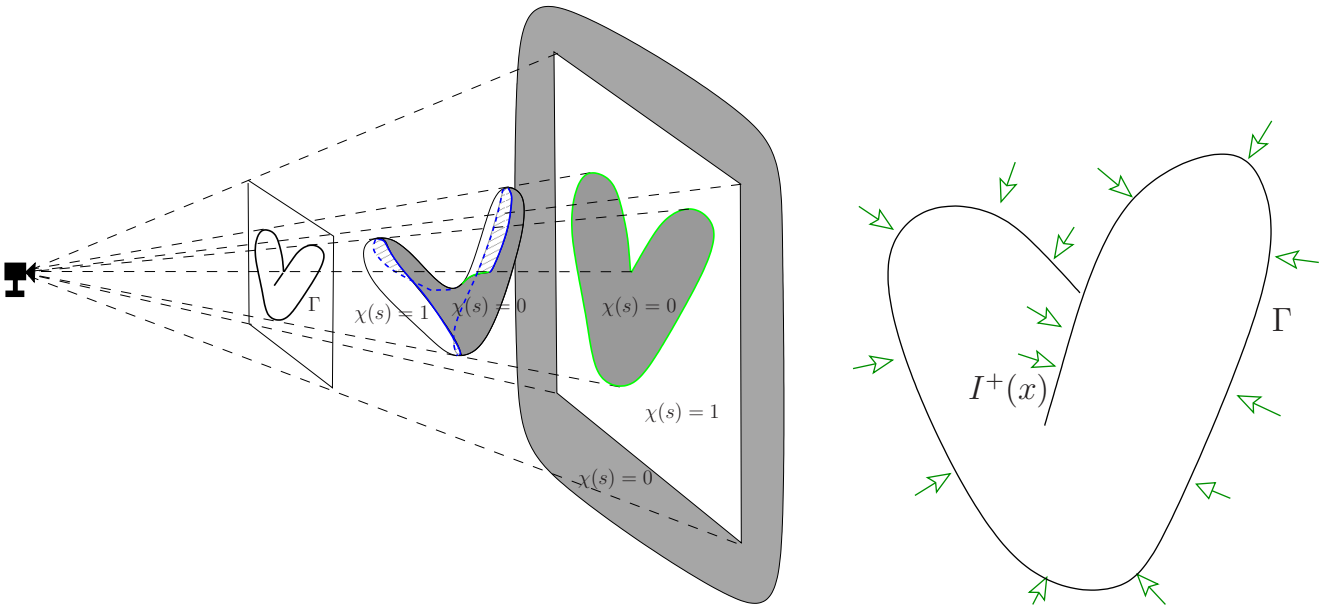
$$v(x, \theta_s) = \frac{\partial \pi_{camera}((1-t)V_a(\theta_s) + tV_b(\theta_s))}{\partial \theta_s}$$
$$= (1-t)\frac{\partial \pi_{camera}(V_a(\theta_s)}{\partial \theta_s} + t\frac{\partial \pi_{camera}(V_b(\theta_s))}{\partial \theta_s}$$

Analyse : définition de l'image I^+_{synth}

Pour pouvoir formaliser les termes du gradient dus aux déplacement des occlusions Γ_{θ_s} dans l'image on définit :

- $\hat{n}_{\Gamma_{\theta_s}}(\mathbf{x})$ la normale à la courbe en $\mathbf{x}$
- On étend l'image par continuité sur Γ_{θ_s} à partir des zone quasi-occludées :

$$I^+_{synth}(\theta_s, \mathbf{x}) = \lim_{k \to \infty} \left(I_{synth}(\theta_s, \mathbf{x} + \hat{n}_{\Gamma_{\theta_s}}(\mathbf{x})/k) \right) \tag{4}$$



Analyse : Forces d'occlusion

On peux maintenant introduire les forces d'occlusion :

$$
\begin{aligned}
f_{oc} :& \Gamma_{\theta_s} \rightarrow \mathbb{R}^2 \\
f_{oc}(\mathbf{x}) =& \Big[\rho\Big(I^{+}_{synth}(\theta_s, \mathbf{x}) - I_{obs}(\mathbf{x})\Big) - \rho\Big(I_{synth}(\theta_s, \mathbf{x}) - I_{obs}(\mathbf{x})\Big)\Big]\hat{n}_{\Gamma_{\theta_s}}(\mathbf{x})
\end{aligned}
\tag{5}
$$

Interpretation

- Si l'image observé au niveau du contour Γ_{θ_s} est plus proche de la couleur à droite que de celle à gauche alors force orientée vers la droite et réciproquement
- grande similarité avec les forces pour les régions actives.

Analyse : calcul du gradient

Le gradient de la mesure d'erreur entre l'observation et l'image synthétisée s'écrit alors :

$$\begin{aligned}\nabla_{\theta_s} E_{dt} = & \int_{\Gamma_{\theta_s}} f_{oc}(\mathbf{x}) v_{\Gamma_{\theta_s}}(\mathbf{x}) d\mathbf{x} \\ & + \int_{\Omega \setminus \Gamma_{\theta_s}} \Big[\mathcal{D}\rho\Big(I_{synth}(\theta_s, \mathbf{x}) - I_{obs}(\mathbf{x}) \Big) \mathcal{D}_{\theta_s} I(\theta_s, \mathbf{x}) \Big] d\mathbf{x}\end{aligned} \tag{6}$$

Interpretation

- Le choix des modèle de texture et d'illumination sont tel que $\mathcal{D}_{\theta_s} I(\theta_s, \mathbf{x})$ est définit partout sauf sur Γ_{θ_s}
- Là où il y a des discontinuités (occlusions), apparaissent des forces d'occlusions.

Analyse : calcul de $\mathcal{D}_{\theta_s} I(\theta_s, \mathbf{x})$

- Le calcul de $\mathcal{D}_{\theta} I(\theta_s, \mathbf{x})$ fastidieux maist 'mécanique' : appliquer la règle de dérivation de fonction composée au processus de synthèse de l'image i.e en combinant les matrices jacobiennes associé chaque étape du calcul de synthèse.
- En combinant les jacobienne à partir de la fin ('mode adjoint' dans le contexte de la différentiation automatique) on ne fait que des produits vecteur-matrice et le coût algorithmique du calcul du gradient peut en théorie être de l'ordre de 4 fois celui du calcul de l'image synthétique
- Cette méthode d'estimation du gradient devient donc plus intéressante que la méthode par différence finie lorsque que le nombre de paramètres dépasse 4

Gradient de E_{dt} par rapport à θ_a et I

De même que précement les gradients $\nabla_{\theta_a} E_{dt}$ et $\nabla_I E_{dt}$ peuvent être obtenus par application 'mécanique ' de la règle de dérivation de fonctions. Par ailleurs, si on a :

- $\rho(x) = \|x\|^2$
- la texture comme fonction linéaire de θ_a (ex PCA)

alors les hessiennes peuvent aussi être calculéé relativement simplement, ce qui pourrais permettre d'accélérer la méthode d'optimisation.

Optimisation

- L'estimation des paramètre de pose de la main(θ_s), de son apparence (θ_a)et des condition d'illuminations(l) se fait par descente quasi-newton du type Broyden-Fletcher- Goldfarb-Shanno (BFGS) en utilisant les gradient $\nabla_{\theta_s} E_{dt}$,$\nabla_{\theta_a} E_{dt}$ et $\nabla_l E_{dt}$.
- La méthode BFGS consiste à construire progressivement une approximation du hessien à partir des gradients obtenus au cours de itérations.
- A chaque étape de la minimisation le dernier gradient obtenu et la hessienne estimée donnent un modèle quadratique de la fonction. La nouvelle pose θ_s est choisie en minimisant cette quadrique
- A chaque nouvelle image, la recherche est initialé par une position prédite à partir des postions estimées précédentes.

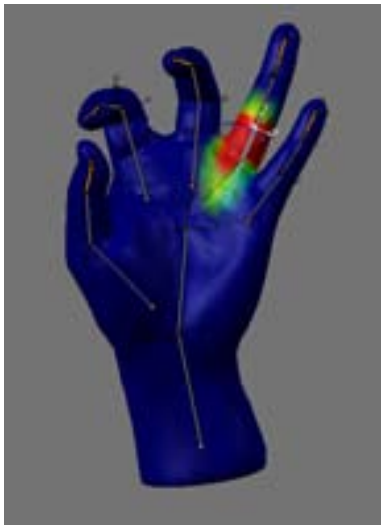
Le suivit de la main

modèle de déformation de la main :

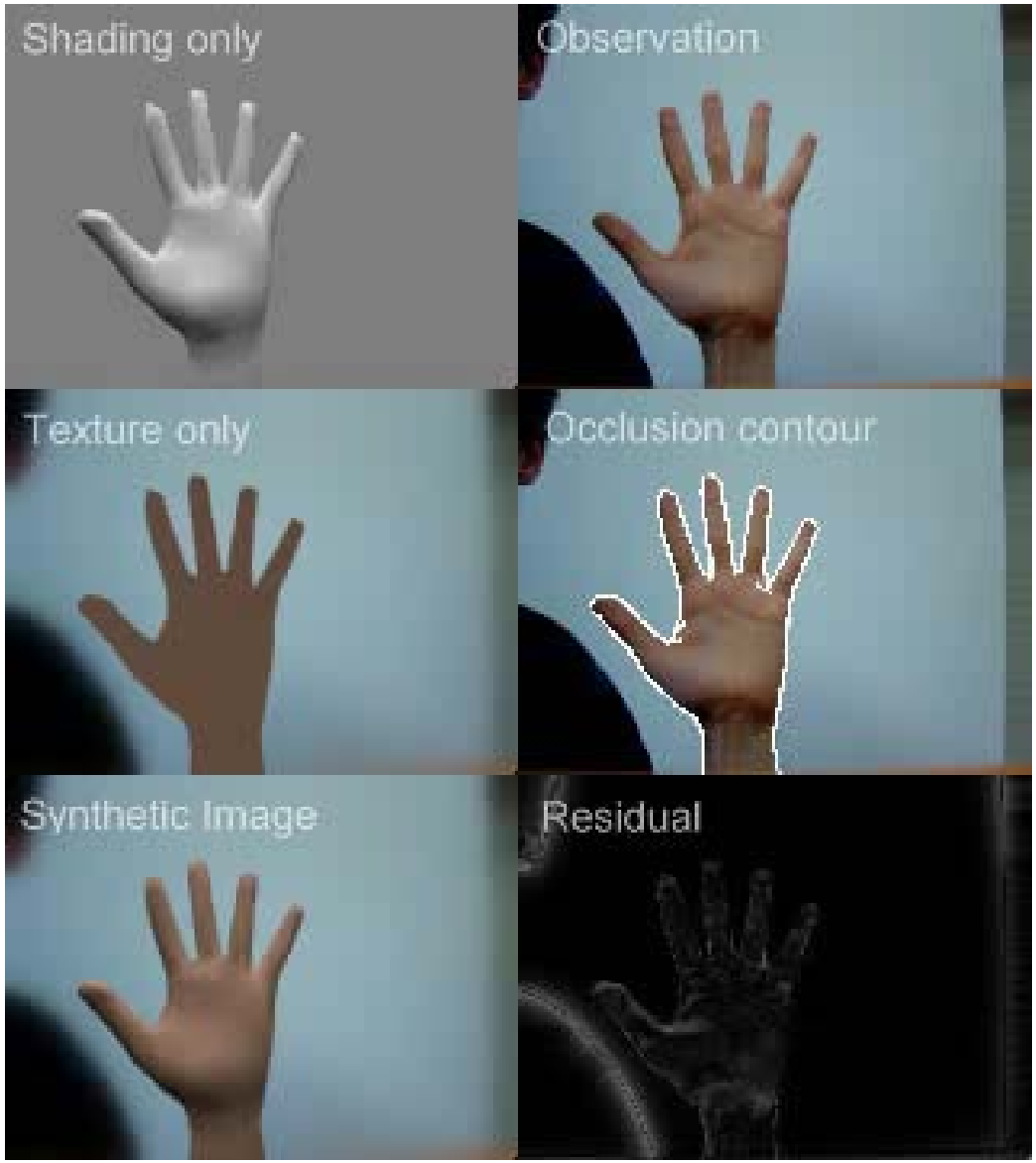
- 17 os et 22 degrées de libertés
- chaque sommet est associé à uneou plusieurs os
- La position d'un somet est données par :

$$V_j(\theta_s) = \sum_i w_{ij} K_{i,\theta_s} K_{i,\theta_{s0}}^{-1} V_j(\theta_{s0}) \tag{7}$$

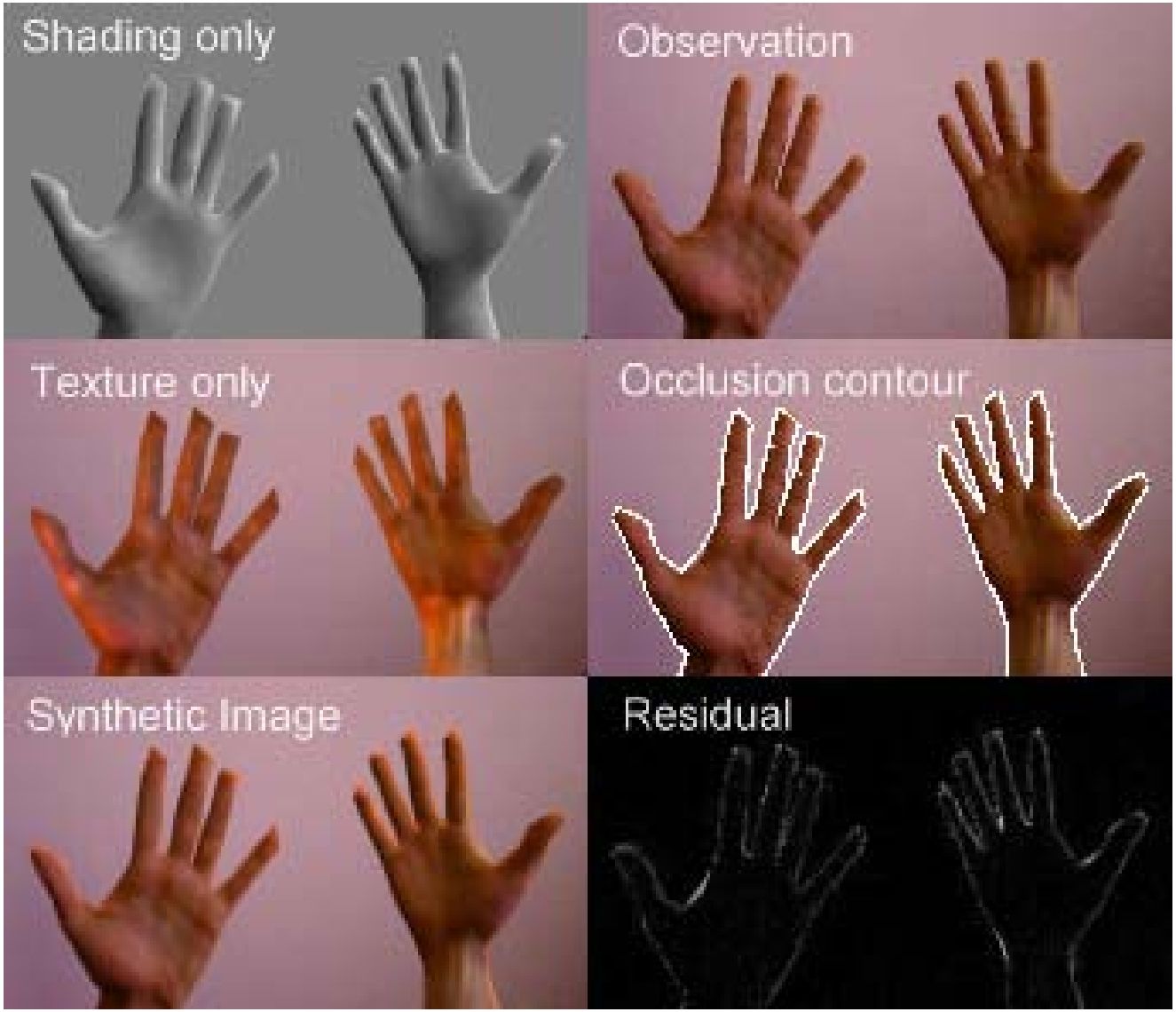
avec K_{i,θ_s} la matrice qui permet de passer des coodonées homogènes dans le repère associé à l'os i aux coordonnées absolues, lorsque que la main est dans la pose θ_s .



Résultats : shading



Résultats : occlusions



Conclusion : Nouveautés

- Calcul du gradient de l'erreur par rapport aux paramètres de la surface dans le cas d'une mesure de l'erreur dans le domaine de l'image
- Traitement rigoureux des variations de l'erreur dues au déplacement des occlusions et auto-occlusions : *forces d'occlusion*
- Convergence vers les contours de la silhouette sans introduire d'autres termes d'attache aux données que la SSD entre image observé et synthétique.
- Utilisation de l'information d'ombrage dans le cadre de suivit de la main

Conclusion : Perspective

- Accélérer l'implementation
- Améliorer l'estimation de la hessienne par rapport à BFGS
- Globaliser la recherche et testant avec différentes initialisations
- Combiner avec des méthode de recherches nonlocales