

Missing data estimation using polynomial kernels

Maxime Berar, Michel Desvignes, Gérard Bailly, Yohan Payan

To cite this version:

Maxime Berar, Michel Desvignes, Gérard Bailly, Yohan Payan. Missing data estimation using polynomial kernels. Lecture Notes in Computer Science, 2005, 3686, pp.390-399. hal-00081903

HAL Id: hal-00081903
https://hal.science/hal-00081903

Submitted on 26 Jun 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Missing data estimation using polynomial Kernels

M. Berar (1), M. Desvignes (1), G. Bailly (2), Y.Payan (3)

(1) Laboratoire des Images et des Signaux, LIS , 961 rue de la houille blanche, BP 46,38402 St Martin d'Hères cedex, France (Maxime.Berar@lis.inpg.fr, Michel.Desvignes@lis.inpg.fr).

(2) Institut de la Communiation Parlée (ICP), UMR CNRS 5009, INPG/U3, 46,av. Félix Viallet, 38031 Grenoble, France (bailly@icp.inpg.fr).

(3) Techniques de l'Imagerie, de la Modélisation et de la Cognition (TIMC), Faculté de Médecine, 38706 La Tronche, France (payan@imag.fr).

Abstract— In this paper, we deal with the problem of partially observed objects. These objects are defined by a set of points and their shape variations are represented by a statistical model. We presents two model in this paper : a linear model based on PCA and a non-linear model based on KPCA. The present work attempts to localize of non visible parts of an object, from the visible part and from the model, using the variability represented by the models. Both are applied to synthesis data and to cephalometric data with good results.

Keywords— PCA, KPCA, statistical models, Image, Pattern recognition.

1 INTRODUCTION

DATA compression, reconstruction, estimation and de-noising are common applications of linear Principal Component Analysis (PCA) [1,2] and Kernel PCA [3,4]. In the latter case, this is a non-trivial task as the results provided by Kernel PCA live in some high dimensional feature space. The main problem of KPCA reconstruction and denoising scheme is to retrieve the data in the input space whose image in Kernel Space is known : in fact, every point of the kernel space does not have a pre image in the input space. This is the pre-image problem [3-5].

In this paper, the estimation of a partially observed object in the input space, using a model learned in the feature space F. is addressed. Some part of the observation is known. To solve this problem, spatial relationships between the known part of the observation and the unknown one are represented in a statistical model and used to localize the unknown part. Those relationships are automatically learned in the model. Like in KPCA reconstruction problem, there are two possible approaches to solve this problem.

The first one use an explicit mapping function φ, the second one use Kernel PCA making φ implicit. In the first case estimation consists in computing the inverse of φ (step 2 in Fig. 1) : a global model (polynomial, sigmoid) of the relations is an a-priori knowledge in this case. In the second case the problem is much more complicate (step 5 in Fig. 1).

The paper is organized as follow : First, the extension of the PCA model to spatial relationship and partial object recognition is presented. Next, the KPCA model is described and the extension to partial object localization is given. Polynomial Kernels are detailed and results are illustrated with synthetic and real examples.

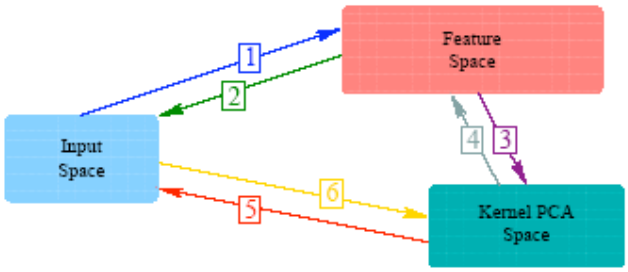


Figure 0: Three different observations space

2 LINEAR PCA MODEL

The extension of the linear PCA model [6] defined here is an elegant way to take into account spatial relations between landmarks and can also estimate the unknown part of the partially visible or occulted model.

Principal Component Analysis is an orthogonal basis transformation, where the new basis is found by diagonalizing the covariance matrix of a dataset.

Let $X_i = (x_{i1}, y_{i1}, ..., x_{in}, y_{in}) \in \mathbb{R}^{2n}$, be the locations of n landmarks. Using PCA, we can write $T \approx \tilde{T} + \Phi b$, where $\tilde{T}$ is the mean shape of the pattern, $\Phi = (\phi_1 | ... | \phi_t)$ is a $(n+m) \times (n+m)$ matrix composed with the eigenvectors of the covariance matrix *S* of the centered data and *b* is a vector of *t* dimension : $b = \Phi'(T_i - \overline{T})$.

The dimension *t* of the vector *b* is the number of eigenvectors with the largest eigenvalues. In classical use PCA, such as de-noising, *t* is chosen by $\sum_{i=1}^{t} \lambda_i \geq 0.95 \sum_{i=1}^{m+n} \lambda_i$. The vector *b* is then a good approximation for the original dataset and any of *n*+*m* points can represented or retrieved with the $t_{t<n+m}$ values of the vector *b*.

Under this hypothesis, if some points (says *t*=*n* points) are known, the remaining unknown points can be determined using PCA. Without any approximations, we can write :

$$\begin{bmatrix} C_1 \\ \vdots \\ C_n \\ X_1 \\ \vdots \\ X_m \end{bmatrix} = \begin{bmatrix} \overline{C}_1 \\ \vdots \\ \overline{C}_n \\ \overline{X}_1 \\ \vdots \\ \overline{X}_m \end{bmatrix} + \begin{bmatrix} \Phi_{1,1} & \cdots & \Phi_{1,n+m} \\ \vdots & \ddots & \vdots \\ \Phi_{n+m,1} & \cdots & \Phi_{n+m,n+m} \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \\ b_{n+1} \\ \vdots \\ b_{n+m} \end{bmatrix}$$

This is a linear system with *n*+*m* equations and unknowns that can not be resolved. Since PCA can represent the dataset with *t*<*n*+*m* values, suppose *t*=*n*, the unknown vector $(b_1, \ldots, b_n, X_1, \ldots, X_m)$ in the following system. Notice, that if we choose t<n, the system become overdetermined and a least square method can be used to resolve the system :

$$\left\| \begin{bmatrix} C_1 - \overline{C}_1 \\ \vdots \\ C_n - \overline{C}_n \\ -\overline{C}_{n+1} \\ \vdots \\ -\overline{C}_{n+m} \end{bmatrix} - \begin{bmatrix} \phi_{1,1} & \cdots & \phi_{1,t} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi_{n,1} & \cdots & \phi_{n,t} & 0 & \cdots & \cdots & 0 \\ \phi_{n+1,1} & \cdots & \phi_{n+1,t} & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ \phi_{n+m,1} & \cdots & \phi_{n+m,t} & 0 & \cdots & 0 & -1 \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_t \\ X_1 \\ \vdots \\ X_n \end{bmatrix} \right\|^2$$

In this framework, a linear approximation of spatial relations between known and unknown points are explicitly determined from the eigenvectors of the covariance matrix.

3 KPCA MODELS

Kernel PCA can be considered as a natural generalization of linear PCA and is very well suited to extract interesting non-linear structures in the data. Closely related to methods applied in Support Vector Machines, it has proved useful for various applications, such as de-noising [] and as a pre-processing step in regressions problems.

3.1 Kernel PCA and Reconstruction

Kernel PCA first map the data from an input space $\mathcal{I}$ into a feature space $\mathcal{F}$ via a (usually non-linear) function and then perform linear PCA on the mapped data. As the feature space $\mathcal{F}$ can be very high dimensional, kernel PCA employs Mercer kernels instead of carrying out the mapping explicitly such as Gaussian kernels $k(x,y) = \exp(-\|x-y\|^2 / c)$ and polynomial kernels $k(x,y) = (1 + x \bullet y)^d$.

Consider data points x and y in the input space $\mathcal{I}= \mathbb{R}^n$. The non-linear mapping $\Phi : \mathbb{R}^n \to \mathcal{F}$ is defined such that :

$$\Phi(x) \bullet \Phi(y) \equiv k(x, y)$$

where $\bullet$ is the vector dot product in the high dimensional feature space $\mathcal{F}$. For a data set $\{ x_i$ i=1 to N$\}$, we have the corresponding set of mapped data points $\{ \Phi_i = \Phi(x_i)$: i =1 to N $\}$in the feature space $\mathcal{F}$. We suppose that our mapped data are centered in $\mathcal{F}$.

To perform PCA in feature space, we need to find Eigenvalues _ > 0 and Eigenvectors $V \in \mathcal{F}\backslash\{0\}$ satisfying $\lambda V = CV$ with $C = \left\langle \Phi(x_i)\Phi(x_i)^T \right\rangle$, the covariance matrix computed on the mapped data. Substituting C into the Eigenvector equation, we note that all solutions V must lie in the span of the Φ-images of the training data. This implies that we can consider the equivalent system:

$$\lambda(\Phi(x_i) \bullet V) = (\Phi(x_i) \bullet CV) \text{ for all i=1 to N} \quad (1)$$

And there exist coefficient $\alpha_1,...,\alpha_n$ such that $V = \sum_{i=1}^{N} \alpha_i \Phi(x_i)$ (2)

Substituting C and (2) into (1), and defining the NxN matrices K (Kernel matrices): $K_{ij} \equiv \Phi(x_i) \bullet \Phi(x_j)$, the problem becomes :

$$\text{solve } N\lambda\alpha = K\alpha \quad (3)$$

To extract non-linear principal components for the Φ-image of a test point $\vec{x}$, we compute the projection onto the k-th component by:

$$\beta_k = (V^k \bullet \Phi(x)) = \sum_{i=1}^{N} \alpha_i^k k(x, x_i) \quad (4)$$

For feature extraction, N kernel functions have to be evaluated instead of a dot product in $\mathcal{F}$, which is expensive if $\mathcal{F}$ is high dimensional (and infinite dimensional for Gaussian kernels). To reconstruct the Φ-image of a vector x from its projections β_k onto the first n principal component in $\mathcal{F}$ (assuming that the Eigenvectors are ordered by decreasing Eigenvalue size), a projection operator P_n is defined by

$$P_n \Phi(x) = \sum_{k=1}^{n} \beta_k V^k \quad (5)$$

When observations are not centered, previous relations are no more satisfied. Observations centering is difficult to achieve in the feature space $\mathcal{F}$, as mapped observations in the feature space and their mean values are not computed for efficiency :

$$\tilde{\Phi}(x) \equiv \Phi(x) - \frac{1}{N} \sum_{i=1}^{N} \Phi(x_i) \quad \forall \vec{x} \in \mathbb{R}^n. \quad (6)$$

In term of dot product, this leads to replace the Kernel matrix by the Gramm matrix. The matrix to diagonalize is then :

$$\tilde{K}_{ij} = K_{ij} - \frac{1}{N} \sum_{p=1}^{N} K_{ip} - \frac{1}{N} \sum_{q=1}^{N} K_{qj} + \frac{1}{N^2} \sum_{p,q=1}^{N} K_{pq} . \quad (7)$$

3.2 Missing Data Estimation

The problem to solve is the reconstruction of partially unknown examples from the KPCA model and from the known part of the data.

Let $z = (c_1,\ldots,c_n,x_1,\ldots,x_m)$ be an example to reconstruct, with the n first coordinates known. The statistical model can be seen as some variability parameters (b in PCA model, β in KPCA model) around a mean shape. Finding the unknown part of x is equivalent to find the shape belonging to the model (*i.e.* variability parameters) whose first coordinates are given by the known part of x. However we are interested in an estimation in the input space ($x_1,x_2,\ldots x_m$) rather than in feature space ($\beta_1,\beta_2,\ldots\beta_k$). So the

solution is given by a vector satisfying $P_n\Phi(c)=\Phi(z)$, which is the pre-image with $(x_1,x_2,\ldots x_m,\beta_1,\beta_2,\ldots\beta_k)$ as unknown. Remember that in the classical pre-image, the feature space coordinates $(\beta_1,\beta_2,\ldots\beta_k)$ are known.

When the vector has no pre-image z, the vector z, such as its image is the nearest one to the model, is found by minimizing

$$\rho(x)=\left\|\Phi(z)-P_N\Phi(c)\right\|^2, \text{ i.e.} \qquad \rho(x)=\left\|\Phi(z)\right\|^2-2\left(\Phi(z)\bullet P_N\Phi(c)\right)+\left\|P_N\Phi(c)\right\|^2 \quad (8)$$

Using equations (5) and (4), kernel notation is introduce to obtain:

$$\rho(x)=k(z,z)-\sum_{k=1}^{N}\left(\sum_{i=1}^{L}\alpha_i^k k(c,x_i)\right)\left(\sum_{i=1}^{L}\alpha_i^k\left(2k(z,x_i)-k(c,x_i)\right)\right) \quad (9)$$

The projection of c and z on the KPCA space are the same :

$$\rho(x)=k(z,z)-\sum_{k=1}^{N}\left(\sum_{i=1}^{L}\alpha_i^k k(c,x_i)\right)^2 \quad (10)$$

This is the general case and minimize ρ(z) depends upon the chosen kernel. This equation can be solve by numerical optimization, but this function presents in general a great number of local minima, sometimes numerically instable. Now, the paper is focused on the polynomial kernels.

3.3 Estimation for polynomial Kernel

Let pose $z=(c_1,\ldots,c_n,x_1,\ldots,x_m)$ as the known part of z is the known part of x. For polynomial kernels, we have to minimize

$$\rho(z)_{\min}=(1+x_c\bullet x_c+z_x\bullet z_x)^d-\sum_{k=1}^{N}\left(\sum_{i=1}^{L}\alpha_i^k(1+x_c\bullet x_{ci}+z_x\bullet x_{xi})^d\right)^2 \quad (11)$$

which is a polynomial of degree 2d with m unknowns. The mapping φ is easily retrieved and is explained using a linear combination of monomial and dot product.

3.3.1 Polynomial degree one

As the observation must be centered in the Feature space $k(x,y)=(x\bullet y)$. The mapping in this case is linear.

$$\rho(x)_{\min}=(c\bullet c+x\bullet x)-\sum_{k=1}^{N}\left(\sum_{i=1}^{L}\alpha_i^k(c\bullet x_{ci}+x\bullet x_{xi})\right)^2=C_{00}+\left\|x\right\|^2-\sum_{k=1}^{N}\left(C_{0k}+C_{1k}\bullet x\right)^2$$
$$\text{where}\quad C_{0k}=\sum_{i=1}^{L}\alpha_i^k(c\bullet x_{ci}),C_0=\left\|c\right\|^2,C_{1k}=\sum_{i=1}^{L}\alpha_i^k x_{xi} \quad (12)$$

For an extremum, the gradient has to vanish: $\nabla_x\rho(x)=2x-\sum_{k=1}^{N}\left(2C_{0k}C_{1k}+2\left(C_{1k}\right)^2x\right)=0$

This lead to a necessary condition for the extremum :

$$x=\frac{\sum_{k=1}^{N}C_{0k}C_{1k}}{1-\sum_{k=1}^{N}\left(C_{1k}\right)^2}=\frac{\sum_{k=1}^{N}\left(\sum_{i=1}^{L}\alpha_i^k\left(c\bullet x_{ci}\right)\right)\left(\sum_{i=1}^{L}\alpha_i^k x_{xi}\right)}{1-\sum_{k=1}^{N}\left(\sum_{i=1}^{L}\alpha_i^k x_{xi}\right)^2} \quad (13)$$

Not surprisingly, this is the classical PCA solution related in §II.

3.3.2 Polynomial degree 2

The mapping φ is given by $\varphi(x)=(x_1,\ldots,x_n,x_1^2,\sqrt{2}x_1x_2,\ldots,\sqrt{2}x_1x_n,x_2^2,\sqrt{2}x_2x_3,\ldots,x_n^2)$

Data in the feature space have to be centered.

$$\rho(x)_{\min} = C_{00} + \left\|\tilde{\varphi}_c(z)\right\|^2 - \sum_{k=1}^{N}\left(C_{0k} + C_{1k} \bullet \tilde{\varphi}_x(z)\right)^2$$

$$\text{with } C_{0k} = \sum_{i=1}^{L} \alpha_i^k (\tilde{\varphi}_c(z) \bullet \tilde{\varphi}_c(x_i)) \text{ and } C_0 = \left\|\tilde{\varphi}_c(z)\right\|^2 \text{ and } C_{1k} = \sum_{i=1}^{L} \alpha_i^k \tilde{\varphi}_x(x_i) \quad (14)$$

For an extremum, the gradient with respect to x has to vanish:

$$\left(\tilde{\varphi}_x(z) \bullet \dot{\tilde{\varphi}}_x(z) - \sum_{k=1}^{N}\left(C_{0k} C_{1k} \bullet \dot{\tilde{\varphi}}_x(z) + \left(C_{1k} \bullet \dot{\tilde{\varphi}}_x(z)\right)\left(C_{1k} \bullet \tilde{\varphi}_x(z)\right)\right)\right)$$ Finding the roots of this polynomial is done by classical numerical method such as newton's one or brent'sone. Note that the solution must be close enough to the mean value of the model and between 3 times the eigenvalue around the mean. This is used as initial value and/or bracketed range.

Finding the solution of the general equation (11) give simultaneously the unknown input space data (unknown part of object) and the variability parameter β of the model.

4 RESULTS

4.1 Synthesis data

In this first experiment, a data set of three points (i.e. six values) is generated (fig 2). Three parameters are needed to perfectly describe these data, i.e. 3 is the theoretical optimal number of variability parameters for PCA and KPCA methods. One point lies on a circle, the two others are constant. Independent Gaussian noise is added to every value.

The PCA and KPCA models are trained on a set of 50 samples. The test set is composed of 200 samples.

In this experiment, the last value of each sample is suppressed and this missing data is estimated by our model.

First, the value of the minimization function (6) in the second degree polynomial is plotted on fig 3. A minima is clearly visible, and the width of this minima is the width of the added Gaussian noise

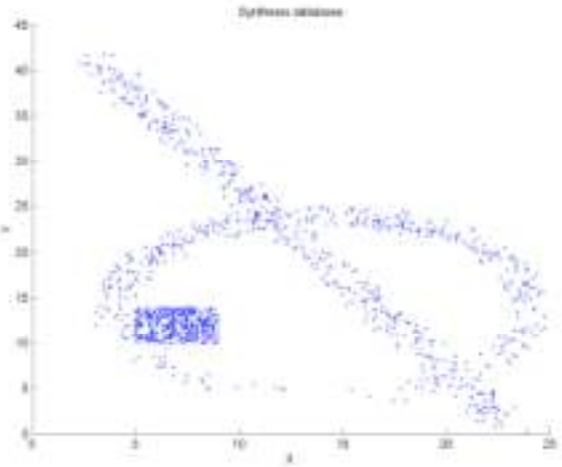


Figure 2 : 3 points with linear and non linear relationships

The error of this estimated unknown value is summarized in the table 1, with respect to the number of variability parameters retained, for 3 methods ::

1. Polynomial Kernel minimization
2. Explicit second degree polynomial projection with PCA : variability parameters are first estimated, following by pre-image computation
3. Classical PCA

Variability parameters	1	2	3	4	5
Kernel minimization	478.94	54.811	60.662	58.685	56.088
Polynomial function	505.1	598.09	592.09	816.96	962.14
PCA	3504.7	27639	19549	39786	4.952e+005

Table 1. : Estimation error for a varying number of parameters

The results exhibit a large advantage to the non linear method : the non linear aspect of the data is well extracted and represented by these models. Linear PCA cannot deal with such non linear data. The second method, in which the variability parameters are first estimated and then the pre-image computed is

less powerfull than the use of the kernel trick and the estimation of the variability parameters and the unknown values in one step.

Another Comparison between linear and Kernel PCA can be achieved with the accuracy of the reconstructed points when the number of these reconstructed points grow. In this example, 3 parameters are needed to described the data. So, 3 values can be retrieved by this method. Figure 4 plots the error of global reconstruction when 1, 2 and 3 points are missing, with number of parameters used on the x-axis. It becomes clear that non linear method has a large advantage, with an increase of computational cost because more parameters are used.

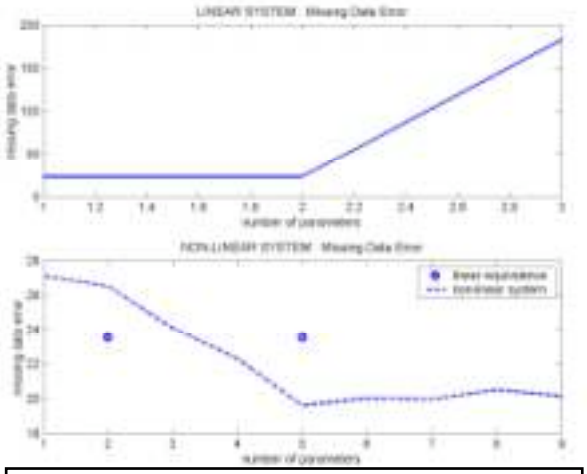


Figure 3 : Reconstruction error for 3 missing points

4.2 Cephalometric data

The goal of cephalometry [2,7] is the study of the skull growth of young children in order to improve orthodontic therapy. It is based on the landmarking of cephalometrics points on tele-radiography, two dimensional X-ray images of the sagittal skull projection. These points are used for the computation of features, such as the length or the angle between lines. The interpretation of these features is used to diagnose the deviation of the patient form from an ideal one. It is also used to evaluate the results of different orthodontic treatment. Cephalometric landmarks are linked to the shape of the cranial contour. In this context, the cranial contour is sampled and the landmark are learned together with the sampled contour [8].

Landmarking a new cephalogram, knowing the contour, is to retrieve unknown part of the model (landmarks), with the model and the known part (sampled contour).

On these real data, linear PCA and KPCA give the same results, with 4mm of mean error. This means that the data are non really non linear, or that the non-linearity cannot be represented by a polynomial of degree 2. This is quite more than a previous non linear and affine invariant version, which use an ad-hoc projection function.



Figure 4: cephalogram, cranial contour and landmarks

5 CONCLUSION

In this paper, a polynonmial kernel based model has been presented. This non linear model is used to resolve the problem of missing data in an image in a statistical framework. We found equation 6, which can be numerically solve in the general case. Shape parameters and missing data are then estimated. With polynomial kernel, we have to found the roots of a polynomial equation and the solution more robust.

The polynomial kernel based model has been compared to classical linear PCA on synthetic and real data. When there is a non linear relationship between data, the kernel model has better accuracy than the linear one, with a larger computational cost.

REFERENCES

[1] T.F.Cootes, G.J. Edwards, C.J.Taylor. Active Appearance Models IEEE PAMI, Vol. 23 (6), pp. 681-685, 2001.

[2] T.J. Hutton, S.Cunningham, P. Hammond. An Evaluation of Active Shape Models for the Automatic Identification of Cephalometric Landmarks. European Journal of Orthodontics, Vol. 22(5), pp. 499-508, 2000.

[3] S. Mika, B. Schölkopf, A.J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch. Kernel PCA and de-noising in feature spaces. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, Advances in Neural Information Processing Systems 11, pages 536-542. MIT Press, 1999.

[4] B. Schölkopf, A. Smola et K. Müller. Non linear component Analysis as a Kernel Eignevalue Problem. Neural Computation, 10(5): 1299-1319, 1998.

[5] J. T. Kwok et Ivor W. Tsang. The Pre-Image Problem in Kernel Methods. Proceedings of ICML 2003 : pp.408-415, 2003.

[6] S. Sclaroff, AP..Pentland, Modal Matching for Correspondence and Recognition. IEEE Transactions on Pattern Recognition and Machine Intelligence, 17(6):545-561, 1995.

[7] B. Romaniuk, M. Desvignes, M. Revenu, M.J. Deshayes Linear and Non-Linear Model for Statistical Localization of Landmarks, ICPR, Vol. 4, pp. 393-396, 2002

[8] 3. B. Romaniuk, M. Desvignes, "Contour Tracking by Minimal Cost Path Approach. Application to Cephalometry", International Conference on Image Processing ICIP, Singapore, October 2004