Virtual, temporally coherent reconstruction of moving animals

Estelle Duveau
Email: estelle.duveau@ensimag.imag.fr

Feb-Jun 2008

M2R Informatique
Spécialité: Image, Vision, Robotique, Réalité et Mondes Virtuels
Parcours: Synthèse et Animation des Mondes Virtuels

Supervisor: Lionel Révéret
Email: lionel.reveret@inrialpes.fr

EVASION/LJK, INRIA Rhône-Alpes - ZIRST
655 Avenue de l'Europe
38334 Saint-Ismier CEDEX
This document is the Master thesis of a Master project done from February to June 2008 in the EVASION team of the LKJ laboratory, in INRIA Rhône-Alpes under the supervision of Lionel Revéret. I would like to thank him as well as the EVASION team for their kindness. I would also like to thank Franck Hetroy for having supervised my project from afar.

This document is organized as follows. Section 1 introduces the project. In section 2, I review the existing techniques. In section 3, the accomplished work is explained. In section 4, the results of this work are discussed and some future work and perspectives are presented.
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Example of 3D measurements of the skeleton of a rat</td>
</tr>
<tr>
<td>2</td>
<td>Example of non-rigid connection between flesh and bones</td>
</tr>
<tr>
<td>3</td>
<td>Long-term goal</td>
</tr>
<tr>
<td>4</td>
<td>Sample of fur taken from the videos of a moving rat</td>
</tr>
<tr>
<td>5</td>
<td>Experimental set-up for [ZSCS04]</td>
</tr>
<tr>
<td>6</td>
<td>5 consecutive frames used in [ZSCS04]</td>
</tr>
<tr>
<td>7</td>
<td>Examples of tracked meshes in [ZSCS04]</td>
</tr>
<tr>
<td>8</td>
<td>Circle of normals that fit the irradiance equation at a given point of the surface</td>
</tr>
<tr>
<td>9</td>
<td>Problem to model our data in order to apply SVD</td>
</tr>
<tr>
<td>10</td>
<td>Results from [HS05] : (a) Input data, (b) Views of the reconstructed model</td>
</tr>
<tr>
<td>11</td>
<td>Edge detection applied to hair in [PBnS04]</td>
</tr>
<tr>
<td>12</td>
<td>Results from [PBnS04]</td>
</tr>
<tr>
<td>13</td>
<td>The shape from shading problem, from [PF06]</td>
</tr>
<tr>
<td>14</td>
<td>Results from [FL93]</td>
</tr>
<tr>
<td>15</td>
<td>Results from [PF06]</td>
</tr>
<tr>
<td>16</td>
<td>Example of relevant shading information on the thigh</td>
</tr>
<tr>
<td>17</td>
<td>Overview of [FRDC04]</td>
</tr>
<tr>
<td>18</td>
<td>External data : 4 cameras seeing the same scene</td>
</tr>
<tr>
<td>19</td>
<td>Internal data : cineradiography seeing the same scene as the cameras</td>
</tr>
<tr>
<td>20</td>
<td>Calibration data used to compute light direction</td>
</tr>
<tr>
<td>21</td>
<td>Recentering videos with respect to marker 5 (red) and 6 (green)</td>
</tr>
<tr>
<td>22</td>
<td>Correlation maps</td>
</tr>
<tr>
<td>23</td>
<td>Correlation maps with another definition of the motion</td>
</tr>
<tr>
<td>24</td>
<td>Data available on hind leg</td>
</tr>
<tr>
<td>25</td>
<td>Patch around marker 9</td>
</tr>
<tr>
<td>26</td>
<td>Variations of the mean intensity of the patch and of the positions of the markers through time</td>
</tr>
<tr>
<td>27</td>
<td>FFT of the mean intensity of the patch and of the positions of the markers</td>
</tr>
<tr>
<td>28</td>
<td>First principal component of PCA on the patch around marker 9</td>
</tr>
<tr>
<td>29</td>
<td>Mean ± 3*principal component (left) and projection on principal components according to time (right) for the first 5 principal components (camera 1 sequence 11h21)</td>
</tr>
<tr>
<td>30</td>
<td>Position of marker 20 with respect to time for sequences 11h13 (left) and 11h21 (right)</td>
</tr>
<tr>
<td>31</td>
<td>Position of marker 16 with respect to time for sequences 11h13 (left) and 11h21 (right)</td>
</tr>
</tbody>
</table>
Position of marker 11 with respect to time for sequences 11h13 (left) and 11h21 (right) .......................................................... 43
Position of marker 14 with respect to time for sequences 11h13 (left) and 11h21 (right) .......................................................... 43
Mean intensity with respect to time for sequences 11h13 (left) and 11h21 (right) .......................................................... 44
Projection on first principal component with respect to time for sequences 11h13 (left) and 11h21 (right) .......................................................... 44
Trajectories on the Xray of the markers of the hind leg during one cycle of run (left) and during the whole sequence 11h13 (right) .......................................................... 46
Key-positions of the hind leg corresponding to the kernel points: magenta = hip, red = pelvis, green = knee, cyan = ankle, blue = toe 47
Input of prediction scheme .......................................................... 48
Examples of superimposition of predicted markers (colored ones) over X-ray and observed markers (white ones) .......................................................... 49
Output of prediction scheme : marker 20 (hip) .......................................................... 50
Output of prediction scheme : marker 16 (knee) .......................................................... 50
Output of prediction scheme : marker 11 (ankle) .......................................................... 51
Output of prediction scheme : marker 14 (toe) .......................................................... 51
Error made by the prediction of the positions of the markers (in pixels) with respect to time (in frames) .......................................................... 52
Mean error made by the prediction of the positions of the markers .......................................................... 53

List of Tables

1 Lead markers .......................................................... 29
2 Correlation of mean intensity with positions of markers .......................................................... 37
3 Percentage of variance of the data in the first principal component .......................................................... 40
4 Mean and standard deviation of the error made by the prediction scheme in pixels and over the spread of the positions in percentage .......................................................... 53
1 Introduction

1.1 Context

Understanding the functions of the skeletal structures of vertebrates in motor control is a key issue in several research domains. In integrative biology, it is necessary to depict anatomy and functional morphology to improve our knowledge of biodiversity. In medicine, clinical research of some malfunctioning needs exact articular diagnostic in normal and transgenic models. In computer sciences, 3D graphics based on natural models is necessary for realistic animation of virtual creatures. All of these approaches share the same crucial need for pertinent three-dimensional and in vivo measurements of postures and motions of the skeleton.

Figure 1: Example of 3D measurements of the skeleton of a rat

However, biomechanical expertise reports that current techniques of geometrical prediction of the skeletal structures from 3D external markers present limitations in accuracy due to the fact that flesh and muscles are not rigidly connected to the bones structure.

Figure 2: Example of non-rigid connection between flesh and bones
In Fig. 2, we can see that a lead marker glued on the skin of the animal (in white) does not have a rigid movement with respect to the lead markers put internally on the bones at the knee and ankle (in red).

1.2 ANR-Kameleon project

The goal of the ANR-KAMELEON project is to process massive internal and external anatomical data for the studying of skeletal structures of vertebrates. Thus, it aims at analyzing and solving the forementioned limitations by combining different sources of anatomical data: X-ray cineradiography for internal structures of the skeleton and videos for the skin surface.

This way, we could learn the association between a 3D surface flow (skin surface) and a 3D skeleton motion. Once this association is learned, we can animate a 3D surface by animating the skeleton and we can diagnose asymmetry or trembling of the skeletal structures due to neuromotric diseases from the 3D skin surface flow. This way, cineradiography (that is not available in all laboratories) wouldn’t be required anymore.

The project gathers multidisciplinary skills of 4 research groups, specialized into complementary aspects of the analysis and interpretation of 3D motion of vertebrates: anatomy (MNHN), biomechanics (LPBEM), clinical research (LNRS), animation and video analysis (INRIA, project leader).

1.3 Project

The long-term objective is thus to predict the movement of the skeletal structures from the external envelope (skin surface). The aim of this Master thesis is to study this external envelope, focusing on its link with the movement. For this, we have acquired video sequences of moving rats.

Indeed, several problems arise. On top of the non-rigidity between internal and
external markers, the non-homogeneous and texture-less fur of the rat (Fig.4) as well as the constrained and restricted experimental set-up with live animals make traditional approaches of computer vision to recover shape from images (stereo-vision, photometric stereo, shape-from-shading) inadequate in our case as we will see in section 2.

Figure 4: Sample of fur taken from the videos of a moving rat

Hence, we focused on statistical approaches on images to establish a link between the variations of pixel intensities on the videos of the rat and the movement of the skeletal structures (section 3.3). This link enabled us to predict the positions of some internal structures according to those pixel intensities (section 3.4). In conclusion (section 4), we will see that we created a new approach to study movement of quadrupeds from videos: movement from shading.
2 Previous work

Computer vision aims at giving capacities of visual analysis to the computer. Modeling scenes is the primary goal. Those scenes are composed of real objects that can be moving or not. Therefore, computer vision tries to model the shapes of these objects and/or model their movement. There are mainly two kinds of approaches: geometric and photometric ones.

2.1 Geometric approaches

Computer vision has for a long time consisted in geometric approaches only. In geometric approaches, shapes and features are considered. The cues taken into account are silhouettes (in multi-view), corresponding features on different cameras (in stereo and multi-view), etc [HZ04]. However, these kinds of reconstruction usually lack temporal coherence or require either a high number of cameras or the use of markers.

Recently, [ZSCS04] proposed a novel approach to model and animate high-resolution faces from videos. Their set-up (Fig.5) consists of 6 synchronized cameras running at 60 frames-per-second (fps) and 2 structured light projectors. The structured light provides high-level features on objects lacking them.

![Figure 5: Experimental set-up for [ZSCS04]](image)

Every 3 frames, no structured light is projected to capture the color texture map of the face and the optical flow.

![Figure 6: 5 consecutive frames used in [ZSCS04]](image)
In this paper, they first compute, for each frame with structured light, a depth map. This map gives for each pixel the distance from the camera i.e. the 3D position of the point. This depth map is computed by finding corresponding points on a pair of stereo images: \( l \) for a camera and \( r \) for the other camera (see [HZ04] for more details). In order to track these correspondences, a disparity function is computed. This disparity function \( d \) gives for each pixel the difference between the pixel coordinates on both images. For example, pixel \((x, y)\) on image \( l \) has coordinates \((x + d(x, y), y)\) on image \( r \) (the images have been rectified). The disparity map is computed by minimizing the error:

\[
E(d(x_0, y_0)) = \sum_{(x,y) \in W_0} e(I_l(x, y), I_r(x + d, y))
\]

where \( W_0 \) is a small window around \((x_0, y_0)\) and \( e \) is a metric between pixels from two cameras.

The contribution of [ZSCS04] is to compute the disparity map on a spacetime window as a global optimization problem to avoid artefacts due to the computation of the disparity at each pixel independently in time and space.

Once this disparity map is computed (hence, the depth maps), to compute a single time-varying mesh, they fit the depth map to a template. The initialization is done with a few corresponding points defined manually on the template and the images and then, a minimization process takes place to fit the first mesh. After that, template tracking using the optical flow creates a single time-varying mesh.

As a result, the output is a 20 fps sequence of high-resolution 3D meshes with no need of markers.

However, these kinds of scheme rely on high-frequency data on the shape. Indeed, traditional stereo works best on rough surfaces or textured regions (even forced to be textured with structured light). Yet, feature points do not exist on fur, numerous markers can’t be used on rodents and structured light can’t be projected fast enough (we use cameras shooting at 200Hz) to capture a rat’s movement.
The only way these approaches could work was if we had stereo pairs of cameras really close to one another so that they could see the very same scene. Then, a comparison of patches of fur could lead to interesting results. However, we will see in section 3.1. that our cameras are at 4 very different viewpoints.

An alternative to using feature points is to use shading information. Indeed, the intensity of a pixel conveys information on the geometry of the surface without requiring correspondences between features.

### 2.2 Photometric approaches

Another approach is using photometric cues to recover shape. Photometric reconstruction aims at recovering geometrical information from shading information. The idea, introduced in [Woo80], comes from the fact that the data we’re working on are pixel intensities depending on the shape of the object.

A link between these intensities and the surface is given by the irradiance equation (assuming a linear relationship between the grey-level image and the image irradiance):

\[
I(x, y) = R(n(x, y))
\]

where \( I \) is the pixel intensity, \((x, y)\) are the pixel coordinates, \( n(x, y) \) the normal vector to the surface projecting into \((x, y)\) and \( R \) the reflectance map.

This equation is usually used in rendering to compute the intensity knowing the normal to the surface. What we want in our case is the reverse process: knowing \( I \) and \( R \), (1) gives us information on \( n \), hence, on the shape. The main advantage over traditional stereo is that it does not require high-level features on different images. It only takes into consideration low-frequency data (pixel intensities).

Different reflectance map models \( R \) exist. They can be classified as :

- **lambertian** :
  \[
  R_L(n) = A\rho \max(0, n.l)
  \]
  where \( A \) is the light intensity, \( \rho \) the albedo of the surface, \( l \) the light direction with both \( n \) and \( l \) being unit vectors. The main advantage of assuming lambertian material is the fact that \( R \) becomes independent of the viewpoint and only one parameter of the surface is involved: the albedo. The albedo is the extent to which an object diffusely reflects light.

- **specular** : several models use a specular component \( R_S \). For instance, in the Phong model :
  \[
  R_S(n) = A_k s(r(n, l).v)\beta
  \]
where $k_s$ is the specular coefficient (the extent to which an object specularly reflects light), $r$ is the direction that a perfectly reflected ray of light would take from this point on the surface and thus depends on $n$ and $l$, $v$ is the direction towards the camera and $\beta$ is the shininess, which decides how evenly light is reflected from a shiny spot.

- hybrid:
  
  \[ R = (1 - \alpha)R_L + \alpha R_S \]

- more sophisticated models.

If we take the simplest model, the lambertian one, we have, for a surface facing the light source:

\[ I(x, y) = A\rho n(x, y).l. \]

This is a linear equation with 3 unknowns: the 3 coordinates of $n$: $n_x, n_y, n_z$:

\[ I(x, y) = A\rho (n_xl_x + n_yl_y + n_zl_z). \]

As a result, there is not one solution only. Indeed, $n$ being a unit vector, a “circle of vectors” fits this equation: these are all the unit vectors whose orthogonal projection on $l$ equals $\frac{I(x, y)}{A\rho}$ (see Fig.8).

![Figure 8: Circle of normals that fit the irradiance equation at a given point of the surface](image)

To solve this ambiguity: two main methods exist: photometric stereo and shape from shading. Photometric stereo uses images of a fixed object from a fixed viewpoint under varying illumination. On the other hand, shape from shading uses one image to compute the normal field.
2.2.1 Photometric stereo

Photometric stereo relies on the shading information given by different illuminations. Indeed, if we have more equations (1) with the same number of unknowns, a system can be solved to find the unknowns $n_x, n_y, n_z$.

Thus, in the case of a lambertian reflectance map, under $f$ independent light directions $(l_1, \ldots, l_f)$, we have $f$ intensities $(I_1, \ldots, I_f)$ at each pixel but only one normal $n$. As a result:

\[
\begin{pmatrix}
I_1 \\
I_2 \\
I_3 \\
\vdots \\
I_f
\end{pmatrix}
= 
\rho
\begin{pmatrix}
l_{1,x} & l_{1,y} & l_{1,z} \\
l_{2,x} & l_{2,y} & l_{2,z} \\
l_{3,x} & l_{3,y} & l_{3,z} \\
\vdots & \vdots & \vdots \\
l_{f,x} & l_{f,y} & l_{f,z}
\end{pmatrix}
\begin{pmatrix}
n_x \\
n_y \\
n_z
\end{pmatrix}
\]

Several approaches exist based on this system. It can be simply used to improve an existing geometric approach by correcting the normals [RB99] or it can be the core of the reconstruction process such as with the Singular Value Decomposition techniques or the example-based ones. Moreover, this can be applied to human hair, which can be considered as close to fur.

**Singular Value Decomposition**:

One interesting technique is the one using the Singular Value Decomposition (SVD). Given a real $m \times n$ matrix $M$, there exists a factorization, the singular value decomposition, of $M$ of the form:

\[
M = U \Sigma V^T
\]

where $U$ is an $m \times m$ orthonormal matrix, $\Sigma$ is an $m \times n$ non-negative diagonal matrix and $V^T$ is the transpose of $V$, an $n \times n$ orthonormal matrix.

In [Hay94, YSEN99, XC05, BJ07], the SVD is applied to

\[
I = 
\begin{pmatrix}
I_{(1,1)} & I_{(1,2)} & I_{(1,3)} & \cdots & I_{(1,p)} \\
I_{(2,1)} & I_{(2,2)} & I_{(2,3)} & \cdots & I_{(2,p)} \\
I_{(3,1)} & I_{(3,2)} & I_{(3,3)} & \cdots & I_{(3,p)} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
I_{(f,1)} & I_{(f,2)} & I_{(f,3)} & \cdots & I_{(f,p)}
\end{pmatrix}
\]

or its transpose for $p$ pixels under $f$ illuminations. Applying SVD to $I$, we have:

\[
I = U \Sigma V^T
\]
Moreover, from equation (2), we have:

\[ I = LS \]  \hfill (4)

where \( L \) describes the \( f \) lighting conditions and \( S \) the \( p \) normals. Therefore, combining equations (3) and (4):

\[ U\Sigma V^T = LS \]  \hfill (5)

In particular, in [BJ07], they use the principle that the set of images produced by a convex, lambertian object under arbitrary lighting can be well approximated by a low-dimensional linear set of images. This set is 4D for a 1st order approximation, 9D for a second order approximation.

For the 4D approximation, each column of \( S \) in equation (4) is \((\rho, \rho m_x, \rho m_y, \rho m_z)^T\). As a result, \( I \) is of rank 4 and we can rewrite equation (5) as:

\[ LS = U\Sigma V^T \simeq (U\sqrt{\Sigma^4c}) \ast (\sqrt{\Sigma^4r}V^T) \]

with \( L \) of size \( f \ast 4 \) and \( S \) of size \( 4 \ast p \) and where \( A^{4c} \) denotes the first 4 columns of \( A \) and \( A^{4r} \) the first 4 rows of \( A \). We can thus write:

\[
\begin{cases}
\tilde{L} = U\sqrt{\Sigma^4c} \\
\tilde{S} = \sqrt{\Sigma^4r}V^T
\end{cases}
\]

However, there is a linear ambiguity since, for any invertible matrix \( A \), \((\tilde{L}A^{-1}, A\tilde{S})\) also is a valid solution. Several constraints such as reflectance known in a few points, constant light magnitude or integrability discussed in [Hay94, XC05, YSEN99, BJ07] aim at solving this ambiguity.

This technique models a fixed object under a varying light. An idea would be to see it the other way around: a moving object under a fixed light, which is what we have in our set-up. If we consider that the surface is locally rigid, what used to be a matrix \( I \) of \( p \) pixels tracked under \( f \) lighting directions becomes a matrix of \( p \) planes or patches tracked on \( f \) frames. However, not only do we need to be able to track those \( p \) planes through time but we also need to be able to assume that the set of patches have rigid movement so that they can be considered lit by same light. Indeed, if the patches, as a set, do not have a rigid movement then we can’t model the lighting condition as one light whose direction is varying through time illuminating a fixed object (see Fig.9).
Another interesting approach is the example-based one. These techniques are based on the orientation-consistency cue: under the right conditions, 2 points with the same surface orientation must have the same or similar appearance in an image. The right conditions are that the points must be made of the same lambertian material and the light must be directional. Therefore, if we learn some associations (Intensity, Orientation) from examples, we can find the orientation of a point from its intensity in images.

The training can be done on a few “frontier points” as in [VFC05]. “Frontier points” are points where an epipolar plane is tangent to the object and therefore has a known normal. This, however, requires an epipolar geometry, hence, two cameras. The training can also be done on example objects as in [HS05].

In this paper, objects of known geometry with similar materials to the ones of the target object are imaged under the same illumination conditions as the target object. For each pixel, the intensities observed for the $f$ illuminations are stored. For each pixel $q$ of the image of the target object, the goal is then to find the pixel $r$ in the training data such that the observations under the different illuminations of both pixels are the closest possible. The orientation at $q$ on the target object is given by the orientation at $r$ on the example object. This kind of techniques really gives impressive results given the simplicity of the method (Fig.10)
This kind of learning scheme can be interesting for the rat. As it is hard to directly extract robust data from the uneven fur, it can be interesting to use a training/predicting scheme. However, as mentioned before, we can’t achieve varying illumination on the moving rat.

Case of hair:

A topic close to our problematic that photometric stereo has tackled is hair, which can be considered as close to fur. The most interesting work on illumination-based reconstruction of hair is the work of Sylvain Paris. Especially, in [PBnS04], an interesting study of the numerous edge detectors is used to extract 2D hair orientation (Fig.11).
From this 2D orientation on the image plane that gives a 3D plane on which the hair segment can lie, the direction of the light source that creates an highlight on that particular hair segment gives the normal vector in that plane, hence, the 3D direction of the hair segment. This technique creates a high-quality physical representation of hair segments (Fig.12).

However, their set-up is out of reach for us as it requires a huge amount of lights and cameras as well as, once again, a fixed head. Yet, the study of edge detectors may be used to find 2D directions of hairs, giving cues to the shape of the rat.

Conclusion:

Photometric stereo gives impressive results on a broad set of objects even on very specific and complex ones such as hair. However, the set-up required can’t be achieved in our case. Shape from shading does not require such a set-up.

2.2.2 Shape from shading

Shape from shading deals with recovering the 3D shape of the surface from one grey-level image taken with a calibrated camera (see Fig.13).

This is done by trying to solve equation (1), which is a first-order partial differential equation (PDE). As mentioned before, finding a unique solution requires additional constraints since the system is, under realistic reflectance map, under-constrained.

We want to solve this equation for a 3D surface. The first question that comes to mind is : what is the surface? What representation should we adopt to solve this equation?
Representation of the surface:
Several representations have been used:

1. *surface gradient* in [BH89, CCDG04, TSY04, AF07, ZYT07]. For instance, for an orthographic camera, we can model the surface as

\[(x, y) \rightarrow (x, y, z(x, y))\]

where \((x, y)\) are the pixel coordinates of a point and \(z(x, y)\) its depth. The normal used in the irradiance equation is then \(n(x, y) = (−\nabla z, 1)\)

2. *object-based* : assuming a pre-defined representation of the surface can constrain the surface and reduce the numbers of unknowns (two per pixel in previous representation). For instance, the surface can be a spline as in [CCDG06] (unknowns : the control parameters of the spline), a mesh as in [FL93] (unknowns : a defined set of vertices’ positions) or a deformable object as in [SM99] (unknowns : nodes).

The choice of the representation usually depends heavily on the method chosen to solve (1).

Techniques:
The techniques to solve equation (1) can for the most part be classified in 4 categories:

1. *Local approaches* : an assumption is made on the local shape of the surface e.g. the surface can be assumed to be locally spherical at each point.

2. *Linear approaches* : the reflectance map \(R\) is linearised in order to solve the irradiance equation (1).
3. **Minimization** : an energy function is minimized. This energy function can be the sum of different energy functions coming from :

- the irradiance equation used in [BH89, FL93, SM99, CCDG06]: the reprojection of the surface must be the input image :
  \[ \int \int (I - R)^2 dxdy \]

- the smoothness of the surface used in [BH89] : the variations of the normals must be the smallest possible :
  \[ \int \int \left( \frac{\partial}{\partial x} \frac{\partial z}{\partial x}(x, y) + \frac{\partial}{\partial y} \frac{\partial z}{\partial y}(x, y) \right) dxdy \]

- its integrability (see in [ZTCS99]) : the surface height is independent of the order of integration :
  \[ \int \int \left( \frac{\partial}{\partial y} \frac{\partial z}{\partial x}(x, y) + \frac{\partial}{\partial x} \frac{\partial z}{\partial y}(x, y) \right) dxdy \]

- the intensity gradient (see in [ZTCS99]) : the gradients of the reprojection and of the input must be the closest possible :
  \[ \int \int \left( (\frac{\partial I}{\partial x} - \frac{\partial R}{\partial x})^2 + (\frac{\partial I}{\partial y} - \frac{\partial R}{\partial y})^2 \right) dxdy \]

The minimization process requires an initialization. For instance, in [FL93], a mesh obtained by traditional stereo is used as an initialization. The energy function is then a sum of 2 terms : one based on traditional stereo equations and the other on the irradiance equation. This way, they avoid the weaknesses of both traditional stereo and shape from shading. Indeed, traditional stereo works best on textured regions but fails on uniform regions. On the other hand, shape from shading fails on textured regions and works best on surfaces with constant albedo.

![Figure 14: Results from [FL93]](image)
4. **Propagation**: depth and/or orientation are known at certain points (usually at singular points, which are the points with the highest pixel intensity, or at occluding boundaries) and propagated on all the surface. These techniques are based on the mathematical theory of Hamilton-Jacobi-Bellman equations and viscosity solutions (see [PF06]). Equation (1) is written as an Hamiltonian. As a result, there is a general formulation for various problems. The main advantage of this technique is that it is based on sound mathematical theories. As a result, the uniqueness of the solution under certain conditions is proven and the method can be applied to more complex models of the scene than those used by other methods. For instance, not only orthographic projection by the camera is assumed: [TSY04, CCDG04, AF07, ZYT07] also tackle perspective projection.

![Figure 15: Results from [PF06]](image)

**Conclusion**: The usual assumptions made by shape from shading are:

- lambertian material
- constant albedo
- shape with no self-shadows nor self-occlusions

In the beginning, orthographic projection and directional light had to be assumed on top of that, which partly explains why shape from shading became unpopular. Indeed, these approximations, combined with others, used to lead to very poor results on real scenes (see [ZTCS99], which provides a good evaluation methodology). To improve those poor results, more sophisticated techniques have been used. Some couple shape from shading with traditional stereo [FL93]. Others get rid of some assumptions as seen with the propagation techniques. To enforce constant albedo, an inpainting algorithm can be used to remove on objects with a dominating albedo the regions with different albedos (eyes in a face, writings on...
paper [PF06]). Consequently, the interest for shape from shading that had died
down due to poor results started back with more accurate models.

Out of the 4 approaches, 2 are of interest to us. Local techniques rely on as-
sumptions on the surface, which limits their applications. Linear approximations
of the reflectance map can be inaccurate. Therefore, minimization or propaga-
tion techniques are bound to give better results for real scenes than the other
approaches. To be specific, minimization approaches are more robust to noise but
require an initialization. On the other hand, propagation approaches are based
on a sound mathematical theory, which widens the kinds of scene model they
can tackle in terms of lighting conditions and projections but are less robust to
noise. As a result, they are usually used in very specific scenes : bent paper, faces
with make-up in a dark environment, etc. A moving rat with non-uniform fur on
varying background is, for the moment, out of reach for these techniques.

2.2.3 Conclusion

The problems specific to our data is :

• fur
• experimental set-up with live animals
• temporal coherence : we are interested in movement and animation.

As a result, photometric stereo can’t be achieved due to the necessity to have a still
scene under varying illumination. As for shape from shading, it can’t be applied
on fur. Thus, up-to-date photometric approaches can not be used easily on our
data. However, it can be seen on videos that the shading gives information on
the movement (Fig.16). The question is therefore : if we can not reconstruct the
shape, can we at least have some information from the shading on the movement
rather than on the shape?

Figure 16: Example of relevant shading information on the thigh
2.3 Statistical approaches

An answer to this question comes from [FRDC04]. The idea on which this paper is based is that variations in binary images encode information on movement. Indeed, applying Principal Component Analysis (PCA) to binary images coming from the segmentation of the video of an animal in a natural environment, they noticed that the PCA encodes variations due to motion only and that the first 2 principal components could be interpreted in terms of motion (flight phase vs. feet on ground and rising vs. descending phase). From this, they were able to generate 3D animation gaits from video data (Fig.17).

Thus, using statistical approaches on our grey-level images may give us information on the motion of the rat. Moreover, as we already have some information on motion from the X-ray images, we may be able to link the shading information to the 2D trajectories of the internal lead markers computed on the X-ray images.
2.4 Problematic

As we can’t adapt our experimental set-up to the existing techniques:

- traditional stereo that requires high-level features, which don’t exist on fur and/or a huge set-up that we want to do without;
- photometric stereo that requires a still object whereas we’re interested in motion;
- and shape from shading that is sensitive to noise and therefore fails on fur, we have to adapt the existing techniques to our set-up. To do this, we have to find what kind of information we can extract from our shading data, using statistical approaches.
3 Contribution

From a restricted set-up (section 3.1.), we acquire data that need to be pre-processed (section 3.2.). From this pre-processed data, we can establish a link between variations of pixel intensities and motion (section 3.3.). In section 3.4., we will see how we can use this association.

3.1 Set-up

Our set-up consists of:

- 4 calibrated grey-level cameras that capture images of 640 * 480 pixels at 200Hz (Fig.18);
- a calibrated cineradiography i.e. a video of X-rays that capture images of 1280 * 1024 pixels at 200Hz (Fig.19);

![External data: 4 cameras seeing the same scene](image)

Figure 18: External data : 4 cameras seeing the same scene
• a treadmill on which a rat is running. We have 2 sequences of the same rat running in the same environment (cameras, Xray, lighting conditions, etc) : one shot at 11:13 that we will call sequence 11h13 and one shot at 11:21 that we will call sequence 11h21. On both sequences, the same pepper-and-salt rat (Fig.18) is running ;

• lead markers placed on the rat. There are two kinds of markers : internal and external ones. Internal ones are placed by surgery inside the rat on the bones (black points in Fig.19). External ones are glued on the skin of the animal (black points in Fig.18). Table 1 gives the numbers that are labelling the markers, their position on the rat (anatomically) and their category (internal or external).
There are 12 external markers. For those markers, if they are visible at a given frame on at least 2 cameras, their 3D positions can be computed (see [HZ04]). As a result, we have a 3D trajectory (with missing data) for each external marker.
There are 17 internal markers. They can only be visible on the cineradiography. As a result, we can only track their 2D positions in pixel coordinates on the X-ray.

This raw data have to be pre-processed in order to be studied.
<table>
<thead>
<tr>
<th>Number</th>
<th>Position</th>
<th>Category</th>
<th>Number</th>
<th>Position</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>eye</td>
<td>external</td>
<td>16</td>
<td>back left knee</td>
<td>internal</td>
</tr>
<tr>
<td>2</td>
<td>head</td>
<td>external</td>
<td>17</td>
<td>front pelvis</td>
<td>internal</td>
</tr>
<tr>
<td>3</td>
<td>between eyes</td>
<td>external</td>
<td>18</td>
<td>back skull</td>
<td>internal</td>
</tr>
<tr>
<td>4</td>
<td>front left foot</td>
<td>external</td>
<td>19</td>
<td>mid pelvis</td>
<td>internal</td>
</tr>
<tr>
<td>5</td>
<td>front center spine</td>
<td>external</td>
<td>20</td>
<td>back left hip</td>
<td>internal</td>
</tr>
<tr>
<td>6</td>
<td>mid center spine</td>
<td>external</td>
<td>21</td>
<td>back pelvis</td>
<td>internal</td>
</tr>
<tr>
<td>7</td>
<td>front left shoulder</td>
<td>external</td>
<td>22</td>
<td>front left knee</td>
<td>internal</td>
</tr>
<tr>
<td>8</td>
<td>back center spine</td>
<td>external</td>
<td>23</td>
<td>front left shoulder</td>
<td>internal</td>
</tr>
<tr>
<td>9</td>
<td>back left hip</td>
<td>external</td>
<td>24</td>
<td>front left shoulder-blade</td>
<td>internal</td>
</tr>
<tr>
<td>10</td>
<td>back left thigh</td>
<td>external</td>
<td>25</td>
<td>front center spine</td>
<td>internal</td>
</tr>
<tr>
<td>11</td>
<td>back left foot</td>
<td>external</td>
<td>26</td>
<td>mid center spine</td>
<td>internal</td>
</tr>
<tr>
<td>12</td>
<td>front left knee</td>
<td>external</td>
<td>27</td>
<td>back center spine</td>
<td>internal</td>
</tr>
<tr>
<td>13</td>
<td>front left foot</td>
<td>internal</td>
<td>28</td>
<td>mid tail</td>
<td>internal</td>
</tr>
<tr>
<td>14</td>
<td>back left foot</td>
<td>internal</td>
<td>29</td>
<td>back tail</td>
<td>internal</td>
</tr>
<tr>
<td>15</td>
<td>front tail</td>
<td>internal</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Lead markers

### 3.2 Pre-processing of data

The first thing to do is to make sure that some consistency exists in the scene i.e. that the lighting conditions stay the same and that the videos are centered on the rat.

#### 3.2.1 Illumination

The first thing to do is to study the lighting conditions as we are studying shading information. Indeed, if the light is not directional i.e. if the light source can’t be considered as a far light source, the pixel intensity at a given point depends on the 3D position of this point. Indeed, assuming a lambertian model:

\[ I = A \rho(n.l) \]

If the light is directional, the light direction at the point \((x, y, z)\) is

\[ l(x, y, z) = \text{cst}, \]

hence, \(I\) depends only on \(n\) i.e. on the shape of the surface. Otherwise,

\[ l(x, y, z) = (LS_x, LS_y, LS_z)^T - (x, y, z)^T \]

where \((LS_x, LS_y, LS_z)\) is the position of the light source. As a result, \(I\) depends on both \(n\) and \((x, y, z)\), hence, on both the shape and position of the surface. In this case, the orientation-consistency cue (see section 2.2.1-SVD methods) can’t be applied: two points with the same surface orientation can have different appearances in an image, depending on their relative 3D positions.
In order to make sure our lighting conditions can be considered as a single directional light, I used calibration images (see Fig.20). To calibrate the cameras, a 3D cube of known geometry with spheres at each corner is used.

![Calibration data used to compute light direction](image)

Knowing the geometry of the 3D cube and the pixel coordinates of the projections of its 8 corners in an image, the 3D positions of the 8 corners can be computed with the POSIT algorithm, available in the OpenCV library.

Knowing the equation of the plane defining the treadmill and the pixel coordinates of the projections of the shadows of the 8 spheres, the 3D positions of the centers of these shadows can also be computed.

Let’s call the 3D position of one corner \( C \) and the 3D position of its shadow \( S \).

The light direction at \( C \) is then \( L = S - C \).

This light direction has been computed for different corners, different frames and with the projections on different cameras. The result is that the mean light direction (normalized) of 32 light directions is:

\[
L = (-0.3440, -0.0048, -0.9375)^T
\]

with standard deviation:

\[
std = (-0.0415, 0.0281, 0.0145)^T.
\]

Given the round-offs and measurement errors in the geometry of the cube, in the pixel coordinates of the centers of the spheres and shadows and the fact that the treadmill is a bit uneven, the standard deviation can be considered as small and, as a result, the light can be considered as directional.

Furthermore, it is interesting to notice that the directionality of the light comes from the use of a retro-projector. Indeed, during a new experiment, we realized
that the directional light was not created by any spot but by the light of the retro-projector. Thus, a retro-projector makes a better and more easily available directional light than a spot-light.

3.2.2 Center videos

We only have 11 markers tracked through time on the surface of the rat. To try and track some more features, we can center the videos on the rat. If this is quite stable, the position in the images of some regions of the rat (such as the body) should not change from one frame to the other and we can therefore track these regions.

Centering videos requires tracking a point through time (a feature point, the centroid of a region, etc). The only such thing we have is the markers. We are therefore going to center the videos on one of the markers.

I therefore centered the videos with respect to all the (visible) markers. One is particularly interesting: centering the videos with respect to the position of
marker 5 gives a far more stable video. Indeed, this marker, located between the shoulder-blades, creates the smallest jittering of the placement of the rat on the video.

In Fig.21, a comparison of the videos centered around marker 5 (in red) and marker 6 (in green, the next one down the spine of the rat) shot by camera 2 shows this fact. An outline of the rat’s shape on the first frame has been done manually (in black). It is then used as a layer over the whole video. It can be seen that, when the rat centered on marker 5 remains quite at the same position, the rat centered on marker 6 has a lateral swinging movement.

This is an interesting fact from a biomecanical point of view but also for computer animation. Indeed, in skeletal animation (animation of a creature by animation of its skeleton), the pelvis is often taken as the root node of the skeleton. That means that all the other nodes (or joints) of the skeleton are defined with respect to this one. What this result shows is that the point between the shoulder-blades could also be used as a root node.

The data having been pre-processed, the problem is now to find what kind of information the pre-processed videos encode.

3.3 Link between pixel intensities and movement

The question is: can we get specific movement information from the pixel intensities? Are the variations of the pixel intensities clearly related to the underlying movement? A statistical study of the images may lead us to establish a link between images and motion. The first thing to study is whether or not the pixel intensities are correlated to the movement.

3.3.1 Correlation Maps

To find out, I computed correlation maps.

A correlation, often measured as a correlation coefficient, indicates the strength and direction of a linear relationship between two random variables. The correlation coefficient of \((X, Y)\) is obtained by dividing the covariance of the two variables \(X\) and \(Y\) by the product of their standard deviations:

\[
\text{corr}(X, Y) = \frac{C_{X,Y}(0)}{\sigma_X \sigma_Y} = \frac{C_{X,Y}(0)}{\sqrt{C_{X,X}(0)C_{Y,Y}(0)}}
\]
\[ \text{corr}(X, Y) = \frac{\int X(t)Y(t)dt}{\sqrt{\int X(t)^2dt \int Y(t)^2dt}}. \]

The range of this coefficient is \([-1; 1]\). Its value is 1 for an increasing linear relationship, \(-1\) for a decreasing linear relationship. In \([-1; 1]\), its value indicates the degree of linear dependence between \(X\) and \(Y\): the closer its absolute value is to 1, the stronger the correlation between \(X\) and \(Y\). If the variables are independent then the correlation is 0, but the converse is not true because the correlation coefficient detects only \textbf{linear} dependencies between two variables.

In our case, we have two variables depending on time i.e. on the frame \(f\):

- the motion : \(M(f)\)
- a pixel intensity for each pixel \(p\) : \(I_p(f)\).

As a result, for each pixel \(p\), the correlation coefficient between the pixel intensity and the motion \(\text{corr}(M, I_p)\) can be computed. In the end, we have a coefficient in \([-1; 1]\) for each pixel of the input video. If we take the absolute value of this coefficient, we can display the strength of the linear dependence between the motion and the pixel intensities as an image: a correlation map.

The scalar representing the motion has been chosen to be the position of the internal marker 20 (on the back hip) with respect to the position of the internal marker 19 (on the pelvis). This way, the motion information is an internal one, not an external one, thus not depending on any “sliding” of the skin and flesh over the bones. The following correlation maps (Fig.22) were computed on the original videos subsampled once. There is linear mapping between \([0; 1]\) and \([\text{blue}; \text{red}]\).

It can be noticed that the zones with the highest correlations (hence the reddest ones) are the same on both camera 1 and 3 showing the profile of the rat (top row of Fig.22) and on both camera 2 and 4 showing the top of the rat (middle row of Fig.22) for sequence 11h13. For camera 1 and 3, the regions at the top of the back leg as well as at the front shoulder are highly correlated to the motion. For camera 2 and 4, the same zones viewed from the top (therefore on both sides of the rat) are detected. Even more interesting, computing the correlation maps for sequence 11h21, we still detect exactly the same zones (bottom row of Fig.22) as for sequence 11h13.

Choosing the position of the internal marker 16 (knee of hind leg) with respect to the position of the internal marker 19 (on the pelvis) or the position of the internal marker 23 (front shoulder) with respect to the position of the internal marker 25 (front center spine), the regions of highest correlations are still the same (Fig.23 : top row : marker 16, bottom row : marker 23).
These maps prove to us that there is relevant and consistent information about the motion in pixel intensities: the relationship even is linear. To figure out what we can extract from this relevant information, I focused on the hind leg.
Indeed, looking at the correlation maps, it can be seen that the top of the hind leg is highly correlated with the motion. On top of that, we have an external marker (marker 9) in this region (left on Fig.24 in white). Therefore, we are able to track more precisely this region. Internally, we have 5 markers of interest (right on Fig.24):

- marker 19 on the pelvis (in red)
- marker 20 on the hip (in magenta)
- marker 16 on the knee (in green)
- marker 11 on the ankle (in cyan) : this one actually is external but it can be considered as internal as there is not many flesh in this part of the anatomy (see in the right image of Fig.24)
- marker 14 on the toe (in blue).

Marker 19 is taken as a reference and the positions of the other internal markers on the X-ray are computed with respect to the position of marker 19 on the X-ray. From now on, when writing about the positions of those markers, it will imply the relative positions of those markers with respect to the position of marker 19 measured on the X-ray. \( U \) is the horizontal coordinate and \( V \) the vertical one.
As the pixel intensity at the marker’s position is that of the marker and not that of the surface of the rat, we are working on a small patch around marker 9 (32 * 32 pixels : Fig.25).

Correlation:

The intensity at the marker’s position is thus the mean intensity of all the pixels of the patch. Fig.26 shows the mean intensity of that patch and the position of markers 20, 16, 11 and 14 varying through time for sequence 11h21.

We can see that both the variations of intensity and position are cyclic. Furthermore, the period of the cycle is the same. Indeed, the middle column of table 2 gives the correlation coefficients of the positions with the mean intensity. Even
Figure 26: Variations of the mean intensity of the patch and of the positions of the markers through time

though the cycle of the intensity is shifted with respect to the ones of the positions, the correlation coefficients are still strong ($> 0.7$), except for the vertical component of marker 14 (toe), which is not linearly correlated to the mean intensity. This is due to the fact that, when the rat is running, he pushes on the ground with his toes. As a result, the toes stay at the same vertical coordinate whereas the rest of the leg keeps moving.

<table>
<thead>
<tr>
<th>Marker</th>
<th>$(corr(U, I), corr(V, I))$</th>
<th>$(corr(F(U), F(I)), corr(F(V), F(I)))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>(-0.7217, -0.3077)</td>
<td>(0.9570, 0.9479)</td>
</tr>
<tr>
<td>16</td>
<td>(0.7475, -0.7890)</td>
<td>(0.9190, 0.9508)</td>
</tr>
<tr>
<td>11</td>
<td>(0.7376, 0.5689)</td>
<td>(0.9506, 0.9202)</td>
</tr>
<tr>
<td>14</td>
<td>(0.7088, -0.0139)</td>
<td>(0.9483, 0.6960)</td>
</tr>
</tbody>
</table>

Table 2: Correlation of mean intensity with positions of markers

To avoid the effect of the shift between cycles, the same calculations have been
done in the spectral domain. Fig. 27 shows the spectra of the mean intensity and of the positions computed with the FFT. There is a peak at the same sample for both the intensity and the positions, which confirms what we had noticed on the spatial domain: the period of both phenomena is the same. The last column of table 2 shows the very strong correlation coefficients between those Fourier transforms (> 0.9), except, once again, for the vertical motion of marker 14 (toe).

![Figure 27: FFT of the mean intensity of the patch and of the positions of the markers](image)

These results prove that there is a very strong relationship between the variations of the intensity on the patch and the positions of the joints of the skeletal structure of the hind leg. We are next going to study where these variations of intensity come from.
To find out what triggers the apparently meaningful variations of the intensity in that unhomogeneous patch (see Fig.25), I applied Principal Component Analysis (PCA) on these variations.

Principal component analysis (PCA) is a vector space transform. It is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. PCA can be used for dimensionality reduction by keeping the characteristics of the data set that contribute most to its variance i.e. by keeping lower-order principal components and ignoring higher-order ones. Such low-order components contain the “most important” aspects of the data. The principal components are found by the SVD of the covariance matrix of the data $XX^T$. Indeed, we want $Y$ such that:

$$Y = V^TX \text{ with } \text{cov}(Y) = YY^T = D$$

(6)

where $\text{cov}(X)$ denotes the covariance matrix of $X$, $D$ is a diagonal matrix and $V$ is an orthogonal matrix. We can rewrite this constraint as:

$$D = V^TXX^TV = V^T\text{cov}(X)V$$

This is the SVD of the covariance matrix of $X$. Thus, $D$ is the diagonal matrix of the eigen-values of the covariance matrix and $V$ is the orthogonal matrix of the eigen-vectors of the covariance matrix. The first principal component is the eigen-vector associated with the largest eigen-value and so on.

In our case, we are applying PCA to the matrix in which each row represents the intensities of the different pixels of the patch at a given frame and each column the intensities of a given pixel through time:

$$I = \begin{pmatrix}
I_{1,1} & I_{1,2} & \cdots & I_{1,p} \\
I_{2,1} & I_{2,2} & \cdots & I_{2,p} \\
\vdots & \vdots & \ddots & \vdots \\
I_{f,1} & I_{f,2} & \cdots & I_{f,p}
\end{pmatrix}$$

for a sequence of $f$ frames and a patch of $p$ pixels.

The first principal component of this matrix gives us the pixels that vary the most. Hence, it gives us the pixels that “drive” the mean intensity of the patch.

As can be seen on table 3, the first principal component concentrates most of the variance ($> 55\%$). As a result, the mean intensity is mostly dependent on the projection of the data on that component.
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Camera</th>
<th>(First eigen-value)/(Sum of eigen-values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11h13</td>
<td>1</td>
<td>56.1520%</td>
</tr>
<tr>
<td>11h13</td>
<td>3</td>
<td>71.6937%</td>
</tr>
<tr>
<td>11h21</td>
<td>1</td>
<td>56.7394%</td>
</tr>
<tr>
<td>11h21</td>
<td>3</td>
<td>64.7756%</td>
</tr>
</tbody>
</table>

Table 3: Percentage of variance of the data in the first principal component

Looking at Fig.28 and left of Fig.29 which respectively represent the first principal component (PC) for different sequences and the mean patch plus or minus three times the principal component for the first 5 components, we can see that the pixels that vary the most importantly are the ones at the bottom left of the patch (the reddest ones). Looking at Fig.25 and 16, we can see that this zone actually corresponds to a region that the thigh of the hind leg crosses. On the other hand, the top right part of the patch corresponds to a stable zone of the rat, one that does not move a lot. Thus, the region of the patch with highest variations, detected by PCA, is the most relevant and interesting one.

Figure 28: First principal component of PCA on the patch around marker 9

Moreover, on the right of Fig.29, we can see the projections of the data on the principal components through time. Indeed, the patch can be modelled, for \( I \) with zero mean, as (Eq.6):

\[
I = VY
\]

i.e. for each frame \( f \): \( \text{patch}(f) = \text{mean patch} + \text{projection on } PC(f)*PC \).

We can notice that the projection on the first principal component is the least noisy, which means this component is the most consistent through time i.e. the variations of its projection through time really is meaningful.

From all these observations, we can consider that the pixel intensities of the patch can be modeled by keeping only the first principal component.
From the study of correlations and PCA, we have established a relevant and clear relationship between the mean intensity on the patch around marker 9 and the trajectories of the internal markers on the hind leg. As a result, a training/prediction scheme is conceivable.

3.4 Prediction of the movement

Still focusing on the hind leg, given the previous results, the training/prediction scheme could work as follows: the intensity of the patch is the input, the position of the markers 20, 16, 11, 14 are the outputs. We have two sequences seen by two cameras, thus, 4 data sets. First of all, we have to make sure that some consistency exists between the data sets of the different sequences.

3.4.1 Consistency of different sequences

Indeed, if the data we want to predict are completely different and inconsistent in the different data sets, the training/prediction scheme will fail. The different data composing our data sets are:
• the output: the positions of the markers: in Fig.30 to 33, the same patterns can be seen in the same range of pixel coordinates. The non-cyclic phases correspond to phases where the rat is still and the straight ones to phases where the markers can’t be seen on the X-ray. Thus, for sequence 11h13, approximately from frame 200 to 300, 400 to 650, before frame 75 and after frame 750, the markers are out of the X-ray frame and from approximately frame 300 and 400, the rat is not moving on the treadmill. For sequence 11h21, approximately from frame 660 to 920 and before frame 250, the markers have moved out of the X-ray frame.

Figure 30: Position of marker 20 with respect to time for sequences 11h13 (left) and 11h21 (right)

Figure 31: Position of marker 16 with respect to time for sequences 11h13 (left) and 11h21 (right)
Figure 32: Position of marker 11 with respect to time for sequences 11h13 (left) and 11h21 (right)

Figure 33: Position of marker 14 with respect to time for sequences 11h13 (left) and 11h21 (right)

- **the input**: the pixel intensities on the patch:
  - **the mean intensity**: in figure 34, we can once again see the same pattern and range for the mean intensity on the patch. Once again, the non-walking phases can be seen as well as the absence of marker 9 in the video frame. The difference of range between camera 1 and camera 3 comes from the differences in the cameras’ settings (exposition).
  - **the PCA**: In Fig.28, it can be seen that for each camera, the same zones of pixels have the highest importance on both sequences. This means that the first principal component of the PCA is invariant to sequences. As a result, it can be concluded that the projection on this first principal component is a good consistent representation of the data. This is reinforced by Fig.35 that shows that the patterns of this projection are the
same on both sequences and very similar to the ones of the mean intensity in Fig.34. The differences in the first principal components between the 2 cameras come from the fact that both cameras have a different viewpoint of the marker 9 thus on the patch around it.

Figure 34: Mean intensity with respect to time for sequences 11h13 (left) and 11h21 (right)

Figure 35: Projection on first principal component with respect to time for sequences 11h13 (left) and 11h21 (right)

From these comparisons, we have emphasized that the first principal component, the patterns of the mean intensity, of the projection on the first principal component, of the X-ray positions of the markers are independent of the sequence. It can be concluded that a training/prediction scheme between cameras and sequences is possible, based on these relevant data.
3.4.2 Prediction

The mean intensity and positions having similar cycles shifted, a regression with radial basis functions seems to be an appropriate training/prediction scheme. If we have a training set of \( n \) observations i.e. \( n \) inputs :

\[
X = (x_1 \ldots x_n)
\]

(where \( x_i \) is a column vector) associated to \( n \) outputs :

\[
Y = \begin{pmatrix}
y_1 \\
\vdots \\
y_n
\end{pmatrix}
\]

(where \( y_i \) is a row vector) and a new input \( x_* \) for which we want to predict the corresponding output \( y_* \), we have :

\[
y_* = x_*^T (XX^T)^{-1} XY.
\]

This is a linear regression. For a regression with radial basis functions, the inputs are mapped to a new space : a feature space. We won’t be manipulating directly \( X \) or \( x_* \) but their projections in this feature space computed as their image by a function \( \Phi \) :

\[
y_* = \Phi(x_*)^T (\Phi(X)\Phi(X)^T)^{-1} \Phi(X) Y.
\]

Usually, \( \Phi \) depends on the distance of the input from reference points, hence :

\[
\Phi(x) = \begin{pmatrix}
\phi(\|x - k_1\|) \\
\vdots \\
\phi(\|x - k_q\|)
\end{pmatrix}
\]

where \( K = (k_1 \ldots k_q) \) are the kernel points and we can choose

\[
\phi(r) = e^{-r^2}.
\]

For more details, see [RW06].

In our case, the input \( x \) is the projection on the first principal component of the PCA on the patch around marker 9 (examples in Fig.35). This is chosen over the mean intensity because we have seen in the previous section that this projection is as relevant and correlated to the motion as the mean intensity over the patch. Moreover, as we have seen (Fig.28) that this principal component represents the variations of intensity of the region of pixels that is interesting in terms of motion, this data is less noisy than the mean intensity that takes into account all the pixels of the patch, even the uninteresting ones.

The output \( y \) is the horizontal \((U)\) and vertical \((V)\) pixel coordinates of the
markers 20, 16, 11 and 14 on the X-ray (with respect to marker 19 as always). However, applying the regression on these inputs and outputs, the results were not satisfying. Indeed, looking at the cycles of the markers created by one cycle of run of the rat (Fig.36), one can notice that the cycles can be decomposed in 2 parts: when the leg is not resting on the treadmill (from point $A$ to point $B$) and when the leg is pushing on the ground (from point $B$ back to point $A$). The trajectory from $A$ to $B$ is not the same as the one from $B$ to $A$. Indeed, when the input $x$ is such that $x = P$ in an ascending phase of the projection (i.e., if the leg is in the air), the output $y = (U_{20}, V_{20}, U_{16}, V_{16}, U_{11}, V_{11}, U_{14}, V_{14})^T$ shouldn’t be the same as if $x = P$ in a descending phase of the projection (i.e., if the foot lies on the treadmill).

![Figure 36: Trajectories on the X-ray of the markers of the hind leg during one cycle of run (left) and during the whole sequence 11h13 (right)](image)

As a result, the input is chosen to be composed of two values ($I = (P, A)^T$):

- $P$: the projection on the first principal component of the PCA on the patch around marker 9 whose range is $[-1; 1]$
- $A$: a boolean specifying if we are in an ascending phase of the projection or not whose range is $\{0, 1\}$

From this, the kernel points correspond to the 4 extrema of the inputs and the medium positions:

$$K = \begin{pmatrix} 1 & 1 & 0 & 0 & -1 & -1 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

where each column represents a kernel point $(P, A)^T$.

These 6 kernel points correspond to key positions in the cycle of run of the rat (Fig.37).
To sum up, our training/prediction scheme is done with the following settings:

- **Input**: projection on first principal component of PCA on patch around marker 9, ascending or descending phase
- **Output**: 2D positions of internal markers of the hind leg on X-ray
- **Prediction**: regression with radial basis functions (6 kernel points).

The results presented here are done for the following configuration:

- **Training data**: The intensity is measured on camera 1 of sequence 11h21. The positions are measured on the X-ray of sequence 11h21.
- **Prediction data**: The input is the intensity measured on camera 3 of sequence 11h13 (see Fig.38). The output is the positions on the X-ray.
- **Evaluation data**: The predicted positions are compared to the positions measured on X-ray of sequence 11h13.
Fig. 39 shows a few key-frames of the predicted sequence. In white are the observed marker positions (the truth) and in color are the predicted positions. Fig. 40 to 43 show the horizontal (U) and vertical (V) predicted positions (in black) compared to the observed ones (in color) with respect to time.

As expected, the prediction is good except for marker 14 (toe: see section 3.3.2.-Correlation). Indeed, when the toes are pushing the ground, the predicted position is completely inaccurate whereas it is quite accurate for the rest of the cycle (Fig. 39).

A mistake of the prediction can be seen at frame 715 for all markers: there is a fork in the prediction even though the observed trajectory is smooth. Looking at fig. 38, we can notice that it comes from a fork in the input intensity.
Figure 39: Examples of superimposition of predicted markers (colored ones) over X-ray and observed markers (white ones)
Figure 40: Output of prediction scheme: marker 20 (hip)

Figure 41: Output of prediction scheme: marker 16 (knee)
Figure 42: Output of prediction scheme: marker 11 (ankle)

Figure 43: Output of prediction scheme: marker 14 (toe)
To quantify those results, I computed the error made by the prediction in pixel units and with respect to the spread of the positions for the different markers. Indeed, the position of the toe has a greater range of values than the one of the hip. As a result, an error of 20 pixels has not the same effect on the hip as on the toe.

Fig. 44 is the graph of the distance in pixels between the prediction and the truth with respect to time for the different markers. Table 4 gives the mean error and standard deviation of those errors both in pixel units and percentage. Fig. 45 displays around a given marker a circle whose radius is the mean error of the prediction of that marker.

These different observations show that the prediction is quite accurate and satisfying for all markers (see measures in percentage in Table 4). However, the fact that the data is more widely spread for marker 14 and 11 makes the visual results less satisfying for marker 11 (ankle) and not satisfying at all for marker 14 (toe): see measures in pixel units (Table 4 and fig. 44). Moreover, the complete inaccuracy of the prediction of the position of marker 14 is only happening during the phase when the rat is pushing on the ground with his toes. Thus, this inaccuracy is short in terms of numbers of frames. As a result, it doesn’t influence the mean error much. It can be concluded that the results are still poor for the toes but satisfying for the other joints.

Figure 44: Error made by the prediction of the positions of the markers (in pixels) with respect to time (in frames)
<table>
<thead>
<tr>
<th>Marker</th>
<th>20 (hip)</th>
<th>16 (knee)</th>
<th>11 (ankle)</th>
<th>14 (toe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean error (in pixels) :</td>
<td>8.4012</td>
<td>16.2562</td>
<td>33.8784</td>
<td>61.2769</td>
</tr>
<tr>
<td>Standard deviation (in pixels) :</td>
<td>4.8342</td>
<td>10.8852</td>
<td>23.4679</td>
<td>38.6562</td>
</tr>
<tr>
<td>Mean error/Spread of data (in %) :</td>
<td>18.8480</td>
<td>14.2494</td>
<td>20.8934</td>
<td>17.5335</td>
</tr>
<tr>
<td>Standard deviation/Spread of data (in%) :</td>
<td>8.0848</td>
<td>5.8178</td>
<td>8.4247</td>
<td>7.3078</td>
</tr>
</tbody>
</table>

Table 4: Mean and standard deviation of the error made by the prediction scheme in pixels and over the spread of the positions in percentage

Thus, a prediction scheme is reliable to predict the 2D positions of the hip, knee and ankle of the hind leg of the rat. This scheme takes as an input a patch of fur at the top of the hind leg (tracked by a marker). From this input on a given camera and sequence and a training on a different camera and sequence, we are able to predict with not much error the 2D positions of the joints of the hind leg.
4 Conclusion

4.1 Results

I’ve presented an approach to deduce 2D skeletal motion from shading, focusing on the trot of the rat.

4.1.1 Contributions

Aside from technical contributions, some less technical results are interesting. First of all, on a completely experimental level, a retro-projector has been proved to create a directional light of really high quality. Moreover, the fact that centering the videos on the point between the shoulderblades stabilizes the movement of the rat is of interest both in skeletal animation and biomechanics.

On a technical level, relevant correlations between the pixel intensities and the skeletal motion have been found on the whole rat. Moreover, studying the case of the hind leg, this strong correlation has enabled us to use a learning scheme to predict the 2D motion of the joints of the hind leg from the pixel intensities of the fur on the hind leg.

4.1.2 Limitations

This prediction is quite accurate except when the input data, which is a projection on the first principal component, is noisy and doesn’t follow the pattern created by a cycle of trot of the rat (see fork in Fig.38). This prediction scheme has been implemented for the hind leg only. This implementation is marker-dependent. Indeed, we need to be able to track a given patch of the fur : the one on the body that the hind leg crosses during the trot. Therefore, even though once the training is done, we do not need the internal markers anymore, we are still dependent on the external ones.

4.2 Perspective

To overcome this dependence on the markers, a mid-term goal is to use the correlation maps to find the pixels whose intensities are relevant. This way, we could use these pixels for the input of the prediction scheme. This scheme could also be extended to other parts of the body, especially to the front leg.

However, this would only work if the correlation maps are as noiseless as possible. For the moment, no pre-processing aside from centering is done on the videos. As a result, high correlations of the background pixel intensities are detected too. Thus, a pre-processing of the videos would improve the relevance of
the correlation maps. The most two important factors are segmentation to avoid any consideration of the background and distortion. Indeed, a line of 15cm, which is approximately the rat’s length, centered on the rat has a varying size on screen depending on the position of the rat on the treadmill. These variations, coming from the distortion by the camera, is of about 20 pixels in images of size 640 * 480. This affects the relevance of the correlation maps.

An interesting direction would also be to study hairs, inspired by [PBN04]. Detecting hairs with such tools as [Ste98] may give us information on the underlying skin surface. They could even be used as features for traditional stereo if we had two very close views of the rat.

A long-term goal is to apply the same kind of scheme in 3D. The training data would be a 3D skeleton embedded in a 3D surface. The prediction scheme would therefore be, given a 3D surface, to find the 3D skeleton. On this 3D skeleton, pathologic parameters would be measured. This would enable the detection of pathologies without requiring a cineradiography or any invasive device.
References


[CCDG06] Frédéric Courteille, Alain Crouzil, Jean-Denis Durou, and Pierre Gurdjos. Reconstruction de spline 3d par shape from shading : spline from shading. Reconnaissance des Formes et Intelligence Artificielle (RFIA), Tours, January 2006.


