Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 2164–2172
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Towards a measure of deformability of shape sequences ☆

Amit K. Roy-Chowdhury *

University of California, Riverside, CA 92521, United States

Received 20 July 2006; received in revised form 25 April 2007
Available online 5 July 2007

Communicated by G. Borgefors

Abstract

In this paper we develop a theory for characterizing how deformable a shape is given a sequence of its observations. We define a term called "deformability index" (DI) for shapes. The novelty of the proposed method lies in its ability to separate the effects of observation noise from the underlying non-rigid deformation process. The DI is computed from the tracked positions of a sequence of deformable shapes, using an affine camera model. Our method assumes that a deformable shape sequence can be represented by a linear combination of rigid basis shapes, where the weights assigned to each basis shape change with time. The tracked points obtained from the 2D shape sequence are transformed to a 3D shape space, whose dimension is then estimated using spectral analysis methods. The dimension of this shape space determines the number of basis shapes needed to represent the shape sequence, which, in turn, determines the deformability index. Our method is different from existing techniques since it is non-iterative, does not require setting an arbitrary threshold and is able to precisely model the effects of noise in the feature positions. Rigid 3D transformations of the shape are taken into account in estimating the DI; however, the method does not require estimation of 3D structure or motion. Experimental results show that the DI is in accordance with our intuitive judgment. Applications and comparative analysis on 3D deformable object modeling are also presented.

Keywords: Shape sequences; Deformations; 3D modeling

1. Introduction

Shape theory is an active area of research in mathematics, statistics and image analysis because of its ability to provide a representation for complex objects and their deformations. Recently, there has been some work on *shape sequence* analysis, including an understanding of the underlying dynamics (Liu and Ahuja, 2004; Soatto and Yezzi, 2002; Vaswani et al., 2003; Veeraraghavan et al., 2005). In 3D computer vision applications, a 2D shape is usually represented by a finite-dimensional linear combination of 3D basis shapes and a camera projection model relating the 3D and 2D representations (Torresani et al., 2001; Brand, 2001, 2005; Xiao et al., 2004). The method has been applied primarily to deformable object modeling.

One of the unanswered questions in studying shape sequences is how to obtain a quantitative understanding of the deformation the shape undergoes. In fact, in order to properly estimate the 3D deformable models, it is important to quantitatively characterize the degree of non-rigidity. While many of the existing methods obtain an estimate of the number of deformation modes from the rank of the matrix of tracked feature points (the measurement matrix), they are accurate only in the noiseless case (Torresani et al., 2001; Brand, 2001, 2005; Xiao et al., 2004). In the noisy case, some of these methods consider the linear subspace of the tracking data that contains most of the variance (e.g., 99% in Brand, 2005). The existing methods do not

☆ This work was partially supported by the National Science Foundation Award ECS-0622176 and the Bourns College of Engineering Initial Complement Fund.
* Tel.: +1 951 827 7886; fax: +1 951 827 2425.
E-mail address: amitrc@ee.ucr.edu
URL: http://www.ee.ucr.edu/~amitrc

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.06.012

relate the estimate of the amount of deformation in a shape sequence to the *statistics of the noise* in the tracked feature points. This can lead to a larger estimate of the deformation modes induced by the noise.

This paper proposes a method for estimating the number of deformation modes taking into account the statistics of the underlying noise in the shape sequence. It does so by deriving a measure, termed as the "deformability index" (DI), of the shape sequence. Estimation of the DI is non-iterative, does not require selecting an arbitrary threshold and can be done before estimation of the 3D structure. Experimental results of the DI estimates are shown using motion capture data as well as real imagery of different human activities. The results show that the DI is in accordance with our intuitive judgment and corroborates certain hypotheses in human movement analysis studies. Applications and comparative analysis in 3D deformable object modeling are presented. We would like to clarify that our proposed method is not a separate strategy for 3D reconstruction of non-rigid objects.

2. Estimating deformability of a shape sequence

2.1. Background material

We hypothesize that each shape sequence can be represented by a linear combination of 3D basis shapes. Mathematically, if we consider the trajectories of P points representing the shape (e.g. landmark points), then the overall configuration of the P points is represented as a linear combination of the basis shapes as

$$\boldsymbol{S} = \sum_{i=1}^{K} l_i \boldsymbol{S}_i, \quad \boldsymbol{S}, \boldsymbol{S}_i \in \Re^{3\times P}, \; l \in \Re. \tag{1}$$

The choice of K is determined by quantifying the deformability of the shape sequence and will be the studied in detail in Section 2.2. We will assume an affine projection model for the camera.

Given F frames of a video sequence with P moving points, we can obtain the trajectories of all these points over the entire video sequence. These P points for the fth frame can be represented in the measurement matrix, $\boldsymbol{W}$, of size $2F \times P$. Under the affine camera model assumption, the P points of a configuration in a frame f, are projected onto 2D image points $(u_{f,i}, v_{f,i})$ as

$$[\boldsymbol{W}_{2F\times P}]_f = \begin{bmatrix} u_{f,1} & \cdots & u_{f,P} \\ v_{f,1} & \cdots & v_{f,P} \end{bmatrix} = \boldsymbol{R}_f \left(\sum_{i=1}^{K} l_{f,i} \boldsymbol{S}_i \right) + \boldsymbol{T}_f, \tag{2}$$

where,

$$\boldsymbol{R}_f = \begin{bmatrix} r_{f1} & r_{f2} & r_{f3} \\ r_{f4} & r_{f5} & r_{f6} \end{bmatrix} \triangleq \begin{bmatrix} \boldsymbol{R}_f^{(1)} \\ \boldsymbol{R}_f^{(2)} \end{bmatrix}. \tag{3}$$

$\boldsymbol{R}_f$ represents the first two rows of the full 3D camera rotation matrix and $\boldsymbol{T}_f$ is the translation matrix consisting of P copies of the camera translation. The translation component can be eliminated by subtracting out the mean of all the 2D points (assuming that we deal with single objects), as in (Tomasi and Kanade, 1992). We can now form the measurement matrix $\boldsymbol{W}$ with the means of each of the rows subtracted.

Using (2) and putting together the tracked features for all the frames, it can be shown that

$$\boldsymbol{W} = \begin{bmatrix} l_{1,1}\boldsymbol{R}_1 & \cdots & l_{1,K}\boldsymbol{R}_1 \\ l_{2,1}\boldsymbol{R}_2 & \cdots & l_{2,K}\boldsymbol{R}_2 \\ \vdots & \vdots & \vdots \\ l_{F,1}\boldsymbol{R}_F & \cdots & l_{F,K}\boldsymbol{R}_F \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_2 \\ \vdots \\ \boldsymbol{S}_K \end{bmatrix} = \boldsymbol{Q}_{2F\times 3K} \cdot \boldsymbol{B}_{3K\times P}, \tag{4}$$

which is of rank $3K$ (Torresani et al., 2001). The matrix $\boldsymbol{Q}$ contains the pose for each frame of the video sequence and the weights $l_{1,1}, \ldots, l_{F,K}$. The matrix $\boldsymbol{B}$ contains the basis shapes in (1). Since $\boldsymbol{W}$ is rank constrained and is the product of the rotation and basis shapes, it is possible to compute these two parameters.

In (Torresani et al., 2001), it was shown that $\boldsymbol{Q}$ and $\boldsymbol{B}$ can be obtained using singular value decomposition (SVD), and retaining the top $3K$ singular values, as $\boldsymbol{W}_{2F\times P} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$ and $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{D}^{\frac{1}{2}}$ and $\boldsymbol{B} = \boldsymbol{D}^{\frac{1}{2}}\boldsymbol{V}^{\mathrm{T}}$. The solution is unique upto an invertible transformation. Methods have been proposed for obtaining an unique solution using the physical constraints of the problem (Torresani et al., 2001; Brand, 2001, 2005; Xiao et al., 2004). This will, however, not affect the derivation of the deformability index (and hence we do not deal with it further), but would be important in estimating the basis shapes for the applications.

2.2. The deformability index (DI)

The above mentioned rank constraint requires knowledge of K in order to estimate the shape and motion parameters. In the noiseless case, this can be estimated from the rank of the measurement matrix (Xiao et al., 2004; Brand, 2005), while in the noisy case a linear subspace of the tracked features that contains most of the variance of the noisy data is considered (Brand, 2005). The dimension of the subspace is selected arbitrarily, without taking into consideration the exact statistics of the noise. Alternatively, minimum error thresholding (Kale et al., 2004) has been employed, but this involves recomputing the entire model if the error threshold is not met. Another related work is (Irani, 1999), where the author proposed projecting the measurement matrix of a *rigid* scene onto the lower dimensional linear subspace obtained in the noiseless case. This can be extended to non-rigid cases, provided we know the proper subspace dimension for every kind of non-rigid object (since it depends on K), which is clearly not practical. Information theoretic measures like geometric AIC or MDL (Kanatani, 2004) are a possibility, but they would require estimation of the model parameters in order to compute the residuals. The method proposed in this paper for estimating the number of deformation modes takes into account the statistics of the noise, and does not require estimation of the 3D model parameters. It is also non-iterative and does not require determination of an arbitrary threshold. In turn, it leads to a quantitative measure of deformability of a shape sequence (Roy-Chowdhury, 2005).

As a shape deforms, the position of the set of points defining the shape changes from one image frame to the next. Defining a deformability index (DI) depends on the ability to obtain a mathematical description of this shape change. The theoretical derivation which follows does precisely this. It proceeds by using the transformation of the point sequence to a shape space (as shown in Section 2.1) and estimating the dimensionality of this shape space. Spectral analysis provides a method for achieving this purpose, the intuitive notion being that spatial deformations in shape must be reflected in the frequency domain. The dimensionality of the shape space will determine the DI. The statistics of the noise in the sequence of feature positions is taken into account in order to correctly estimate the DI. This is important since the noise can randomly alter the positions of the points, giving a false notion of increased variability in the shape sequence, leading to a higher dimensionality of the shape space. Also, rigid 3D transformations of the shape can provide the impression of deformation. This will be factored out in estimating the DI. However, estimation of 3D structure will not be required for this purpose.

2.2.1. Computation of deformability index

Consider the set of coordinates representing the shape of the deformable object in a particular frame of a video sequence to be the realization of a random process. The sequence of frames depicts the deformation of the shape, along with the effects of the 3D translation and rotation. Represent the u and v coordinates of the sampled points in a *single* frame as a vector $\boldsymbol{y} = [u_1, \ldots, u_P, v_1, \ldots, v_P]^{\mathrm{T}}$. Then, from (4), it is easy to show that for K basis shapes (K is unknown)

$$\boldsymbol{y}^{\mathrm{T}} = \left[l_1\boldsymbol{R}^{(1)}, \ldots, l_K\boldsymbol{R}^{(1)}, l_1\boldsymbol{R}^{(2)}, \ldots, l_K\boldsymbol{R}^{(2)}\right] * \begin{bmatrix} \boldsymbol{S}_1 & \\ \vdots & 0 \\ \boldsymbol{S}_k & \\ & \boldsymbol{S}_1 \\ 0 & \vdots \\ & \boldsymbol{S}_k \end{bmatrix} + \eta^{\mathrm{T}}, \tag{5}$$

$$\text{i.e.,} \quad \boldsymbol{y} = (\boldsymbol{q}_{1\times 6K}\boldsymbol{b}_{6K\times 2P})^{\mathrm{T}} + \eta = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{q}^{\mathrm{T}} + \eta, \tag{6}$$

where η represents the noise in the sequence of tracked points and is assumed to be a zero-mean random process. The vector $\boldsymbol{q}$ is obtained by juxtaposing two consecutive rows of $\boldsymbol{Q}$, corresponding to the same image frame, in Eq. (4). The matrix $\boldsymbol{b}$, which is constant across all the frames, is obtained by duplicating $\boldsymbol{B}$ in Eq. (4).

Assuming that the coordinates of the points representing the shape in all the F frames can be considered to be realizations of the same random process (which is a reasonable assumption since they represent the same shape), with possibly different noise statistics, we can compute the correlation matrix of $\boldsymbol{y}$. Let $\boldsymbol{R}_y = E[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}]$ be the correlation matrix of $\boldsymbol{y}$ and $\boldsymbol{C}_\eta$ the covariance matrix of η. Hence,

$$\boldsymbol{R}_y = \boldsymbol{b}^{\mathrm{T}}E[\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}]\boldsymbol{b} + \boldsymbol{C}_\eta. \tag{7}$$

The correlation matrix, $\boldsymbol{R}_y$, is of size $2P \times 2P$ and can be estimated from the sequence of points representing the shapes as $\boldsymbol{R}_y = \frac{1}{F}\sum_{f=1}^{F}\boldsymbol{y}_f\boldsymbol{y}_f^T$, where $\boldsymbol{y}_f$ is the vector $\boldsymbol{y}$ (defined above) in the frame f. The expectation on the right hand side of Eq. (7) can be computed similarly as $E[\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}] = \frac{1}{F}\sum_{f=1}^{F}\boldsymbol{q}_f^T\boldsymbol{q}_f$, where $\boldsymbol{q}_f$ is the vector $\boldsymbol{q}$ (defined above) for frame f and is obtained from the matrix $\boldsymbol{Q}$ in Eq. (4).

The noise covariance matrix, C_η, represents the accuracy with which the feature points are tracked and needs to be estimated from the image frames. Since η need not be an independent and identically distributed (IID) noise process, C_η will not necessarily have a diagonal structure (but it is symmetric and positive semi-definite). For the purposes of setting a precise threshold (which will become clear soon), it is desirable that $\boldsymbol{C}_\eta$ be a diagonal matrix.

Consider the diagonalization of $\boldsymbol{C}_\eta = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{\mathrm{T}}$, where $\boldsymbol{\Lambda} = \mathrm{diag}[\boldsymbol{\Lambda}_s, 0]$ and $\boldsymbol{\Lambda}_s$ is an $L \times L$ matrix of non-zero singular values of $\boldsymbol{\Lambda}$. Let $\boldsymbol{P}_s$ denote the orthonormal columns of $\boldsymbol{P}$ corresponding to the non-zero singular values. Therefore,

$$\boldsymbol{C}_\eta = \boldsymbol{P}_s\boldsymbol{\Lambda}_s\boldsymbol{P}_s^{\mathrm{T}}. \tag{8}$$

Pre-multiplying Eq. (6) by $(\boldsymbol{\Lambda}_s^{-\frac{1}{2}}\boldsymbol{P}_s^{\mathrm{T}})$, we see that (6) becomes

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{q}^{\mathrm{T}} + \tilde{\eta}, \tag{9}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{\Lambda}_s^{-\frac{1}{2}}\boldsymbol{P}_s^{\mathrm{T}}\boldsymbol{y}$ is a $L \times 1$ vector, $\tilde{\boldsymbol{b}}^{\mathrm{T}} = \boldsymbol{\Lambda}_s^{-\frac{1}{2}}\boldsymbol{P}_s^{\mathrm{T}}\boldsymbol{b}^{\mathrm{T}}$ is a $L \times 6K$ matrix and $\tilde{\eta} = \boldsymbol{\Lambda}_s^{-\frac{1}{2}}\boldsymbol{P}_s^{\mathrm{T}}\eta$. It can be easily verified that the covariance of $\tilde{\eta}$ is an identity matrix $\boldsymbol{I}_{L\times L}$. This is known as the process of "whitening", whereby the noise process is transformed to be IID (Stoica and Moses, 1997).

Representing by $\boldsymbol{R}_{\tilde{y}}$ the correlation matrix of $\tilde{\boldsymbol{y}}$, it can be seen that

$$\boldsymbol{R}_{\tilde{y}} = \tilde{\boldsymbol{b}}^{\mathrm{T}}E[\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}]\tilde{\boldsymbol{b}} + \boldsymbol{I} = \boldsymbol{\Delta} + \boldsymbol{I}, \tag{10}$$

where, for simplicity, $\boldsymbol{\Delta} \triangleq \tilde{\boldsymbol{b}}^{\mathrm{T}}E[\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}]\tilde{\boldsymbol{b}}$. Now, $\boldsymbol{R}_{\tilde{y}}$ is of dimension $L \times L$, $\tilde{\boldsymbol{b}}^{\mathrm{T}}$ is of rank $L \times 6K$ and $E[\boldsymbol{q}^{\mathrm{T}}\boldsymbol{q}]$ is of rank $6K \times 6K$. Thus, $\boldsymbol{\Delta}$ has maximum rank $6K$, where K is the number of basis shapes (assuming $L > 6K$, which is true when there are enough tracked points on the object). This is based on the fact that if $\boldsymbol{A}_{m\times n} = \boldsymbol{F}_{m\times r}\boldsymbol{G}_{r\times n}$, then the $\mathrm{Rank}(\boldsymbol{A}) \leqslant r$. For a general 3D scene undergoing translation and rotation, the rank will be $6K$, which is the case we will consider below. Representing by $\mu_i(\boldsymbol{A})$ the ith eigenvalue of the matrix $\boldsymbol{A}$, we see that

$$\mu_i(\boldsymbol{R}_{\tilde{y}}) = \mu_i(\boldsymbol{\Delta}) + 1, \quad \text{for } i = 1, \ldots, 6K, \quad \text{and}$$
$$\mu_i(\boldsymbol{R}_{\tilde{y}}) = 1, \quad \text{for } i = 6K+1, \ldots, L. \tag{11}$$

Hence, there are $6K$ eigenvalues above 1. By counting the number of eigenvalues that are greater than 1 and dividing it by 6, we can obtain an estimate of K, which is the dimensionality of the shape space represented by the sequence of deforming points. Since K denotes the number of basis shapes that can model the feature point sequence, it provides a measure of the deformability of the shape sequence. The more the number of basis shapes required to model a shape sequence, the more deformable it is. Thus, for a general 3D scene undergoing translation and rotation, we have

$$\text{Deformability Index (DI)} = \frac{\#\text{ eigenvalues of } \boldsymbol{R}_{\tilde{y}} > 1}{6}. \tag{12}$$

The DI is not necessarily a whole number. The number of basis shapes, K, is estimated by rounding the DI to the nearest integer. The division by 6 is because there are $6K$ eigenvalues greater than 1, which is due to the construction of the matrices in (6).

2.2.2. Properties of the deformability index

- For the case of a 3D rigid body, the DI is 1. In this case, the only variation in the values of the vector $\boldsymbol{y}$ from one image frame to the next is due to the global rigid translation and rotation of the object. The rank of the matrix $\boldsymbol{\Delta}$ will be 6 and the deformability index will be 1.
- Estimation of the DI does not require explicit computation of the 3D structure and motion in Eq. (4), since we only need to compute the eigenvalues of the covariance matrix of the 2D feature positions.
- The computation of the DI takes into account any rigid 3D translation and rotation of the object (as recoverable under a scaled orthographic camera projection model). Thus it is more general than a method that considers purely 2D image plane motion.
- The "whitening" procedure described above enables us to choose a *fixed* threshold of one for comparing the eigenvalues.
- The proposed algorithm in non-iterative unlike in some other approaches (Kale et al., 2004; Torresani et al., 2001) where the sum of squared differences is minimized to determine K.
- As evident in the derivation above, the DI is dependent only on the covariance of the measurement data and relies on the rank properties of the structure and motion matrices (not their actual values). Hence it is not affected by the orthonormality constraints required to obtain a unique solution and can be integrated with the methods in (Torresani et al., 2001; Xiao et al., 2004; Brand, 2005).
- For the special case of a planar scene, the corresponding rank of $\boldsymbol{\Delta}$ would be $4K$, and thus the DI should be calculated by dividing the number of eigenvalues over 1 by 4.
- Accuracy of the DI estimates will depend upon correctly estimating the noise covariance matrix, which has been studied extensively by a number of researchers (Trucco and Verri, 1998; Kanatani, 1996; Sun et al., 2001; Daniilidis and Spetsakis, 1993; Weng et al., 1987).

3. Experimental results

We first show experimental results for estimating the DI on motion capture data of different activities, and on human gait analysis data. Thereafter we show experiments in 3D non-rigid modeling using the DI measure. The DI measure has also been used by other authors in non-rigid face modeling (Del Bue et al., 2005).

3.1. Estimating DI on motion capture data

In this experiment, we computed the DI of the human body for a large number of activities and found them to be very consistent with what would be expected intuitively by a human observer. We used the motion capture data available from Credo Interactive Inc. and Carnegie Mellon University. The dataset included a number of subjects performing various activities, like walking, jogging, sitting, crawling, brooming, etc. Some examples are shown in Fig. 1. For each of these activities, we had multiple video sequences consisting of tracked feature points using motion capture data. Also, many of the activities contained video from different viewpoints.

For the different activities in this database, we computed the DI from Eq. (12), as shown in Table 1. We used all the frames in one cycle for cyclic activities, while for the other activities, like brooming, sitting, all the available frames were used. Because motion capture data (which can be quite accurate) was available, we assumed that the noise is small with a standard deviation of 2 pixels. Since the

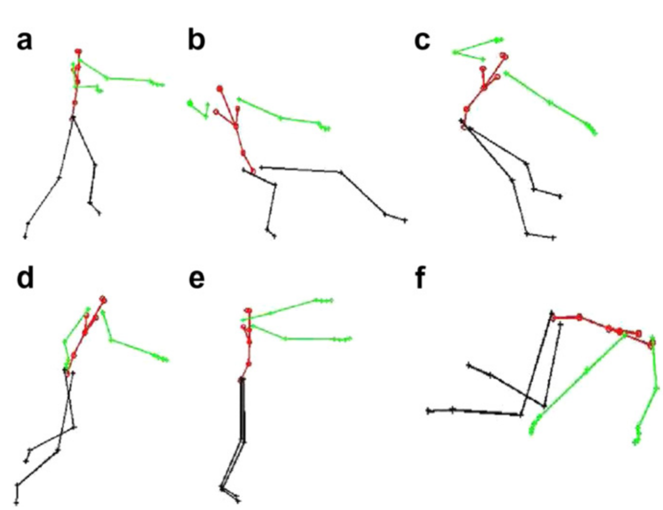

Fig. 1. Plots of the first basis shape, S_1, for walk, sit, broom, jog, blind walk and crawl sequences, respectively.

Table 1
Deformability index for human activities using motion capture data

	Activity	Deformability index
1	Male Walk (Seq. 1)	5.8
2	Male Walk (Seq. 2)	4.7
3	Fast Walk	8.0
4	Walk while throwing hands around	6.8
5	Walk with drooping head	8.8
6	Sit (Seq. 1)	8.0
7	Sit (Seq. 2)	8.2
8	Sit (Seq. 3)	8.2
9	Broom (Seq. 1)	7.5
10	Broom (Seq. 2)	8.8
11	Jog	5.0
12	Blind walk	8.8
13	Crawl	8.0
14	Jog while taking U-turn (Seq. 1)	4.8
15	Jog while taking U-turn (Seq. 1)	5.0
16	Broom in a circle	9.0
17	Female Walk	7.0
18	Slow Dance	8.0

DI is related to the number of basis shapes required to represent the video sequences, we resynthesized the original sequences using the basis shapes and combination coefficients obtained from Eq. (4). Eq. (2) was used for the synthesis and the value of K was determined by the procedure in Section 2.2.1.

From Table 1, a number of interesting observations can be made. For the walk sequences, the DI was between 5 and 6. This matches the hypotheses in papers on gait recognition where it is mentioned that about five exemplars are necessary to represent a full cycle of gait (Kale et al., 2004; Lee and Grimson, 2002). The number of basis shapes increases for fast walk, as expected from some of the results in (Tanawongsuwan and Bobick, 2002). When the person walks doing some other things (like moving head or hands or a blind person's walk), the number of basis shapes needed to represent it (i.e. the deformability index) increases from that of normal walk. Modeling female walk needs more basis shapes than male walk, and this is justified from various studies that have shown that gait cycles of women usually have higher frequency content (Kale et al., 2004; Lee and Grimson, 2002). The result that might seem surprising initially is the high DI for sitting sequences. On closer examination though, it was found that the person, while sitting, was making all kinds of random gestures as in talking to someone else. That increased the DI for these sequences. Also, the DI is insensitive to changes in viewpoint (azimuth angle variation only), as can be seen by comparing the jog sequences (14 and 15 with 11) and broom sequences (16 with 9 and 10). This is not surprising since we do not expect the deformation of the human body to change due to rotation about the vertical axis. Note that DI by itself will usually not be enough to distinguish between activities; however, it is very helpful for identifying models that can be used for activity recognition.

3.2. Estimating DI on Gait Dataset

The USF Gait Challenge Dataset (Sarkar et al., 2005) was used for our experiments because of two reasons. It has a number of examples of different people walking under different conditions. Thus it would allow us to test the consistency of the estimates for the DI. Secondly, a number of researchers have reported results in this dataset and thus we would be able to corroborate our conclusions with their results.

We used the background subtracted images of the walking person, when the person is presenting a side view to the camera. The outer boundary of the person was sampled in order to obtain the shape vector. The method described in (Sun et al., 2001) was adopted to estimate the variance of the noise in the feature positions from the original images. The method uses the inverse of the Hessian matrix of the second-order partial derivatives of the intensity along the horizontal and vertical axes. By using the same number of sample points in each frame, an approximate correspondence was maintained between the feature points in the dif-

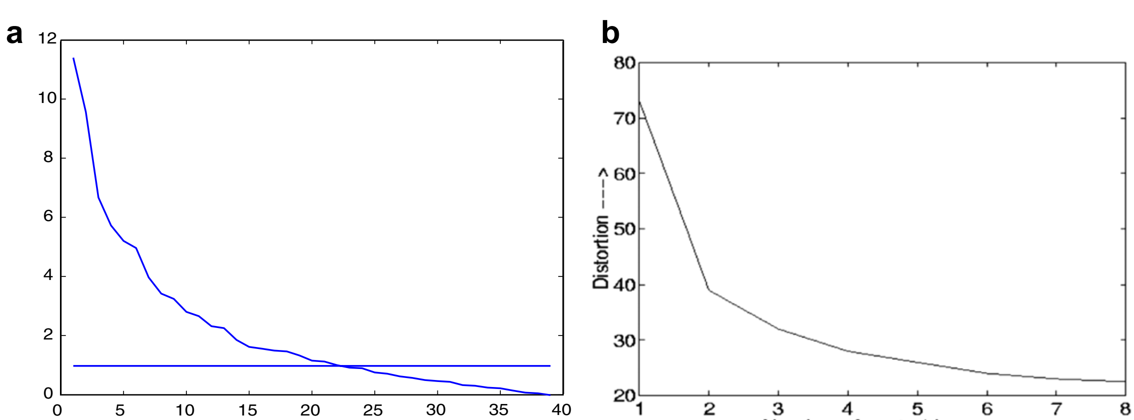

Fig. 2. (a) Plot of the eigenvalues, in decreasing order of magnitude, for a typical walking sequence in the USF database. (b) Distortion vs. number of exemplars on USF gait data (reproduced from Fig. 4 of Kale et al., 2004).

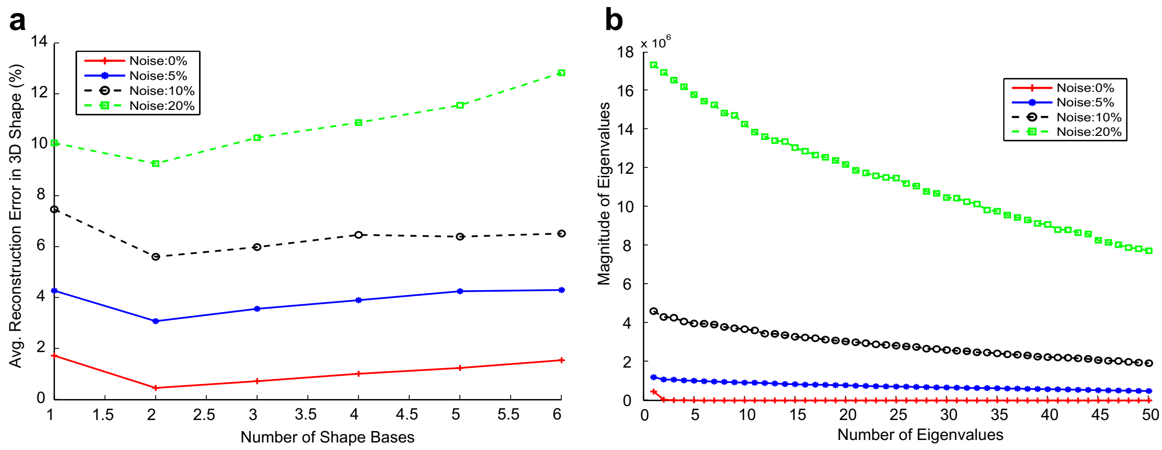

Fig. 3. (a) Average reconstruction error vs. number of basis shapes for the shark sequence in (Torresani et al., 2003). (b) Plot of the eigenvalues of the measurement covariance matrix.

ferent frames. We experimented with 10 subjects walking on grass and concrete surfaces and wearing different types of shoes. For all the cases, the DI ranged from 3.8 to 5.2. Fig. 2a shows a typical plot of the eigenvalues arranged in descending order of magnitude along with the threshold of one (there are some eigenvalues below one due to errors in the estimation of the noise covariance matrix). It has been noted in (Kale et al., 2004) that four to five exemplars are needed to represent a complete cycle of gait. However, this was arrived at using a minimum error thresholding method after repeating the recognition experiment with different numbers of exemplars (see Fig. 4 of Kale et al. (2004), reproduced here in Fig. 2b). Our analysis provides a one-shot, non-iterative estimate and a theoretical justification for the choice of the number of exemplars.

3.3. Estimating DI for 3D deformable modeling

We now show results on modeling 3D deformable objects using the estimate of the number of basis shapes. We also compare our approach against other methods for estimating the deformation modes. We use the 3D modeling approach described in (Torresani et al., 2003) and use the associated software available at http://www-cs-students.stanford.edu/~ltorresa/software.html. We first show the importance of estimating the DI accurately on the synthetic shark data available on this website. We then show some examples on real data.

On the shark data, we computed the DI at various noise levels and using a different number of basis shapes. The noise was added as percentage of the norm of the measurement matrix, $\boldsymbol{W}$. The results are shown in Fig. 3a. From here we see that the minimum error in 3D reconstruction occurred when the number of basis shapes was equal to 2 (which was used in generating this data). Our estimate of the DI was correct (i.e., $K = 2$), which is to be expected as the noise distribution is known. On the other hand, we see from Fig. 3b, that it is very difficult to set a threshold on the eigenvalues of the covariance matrix of the original measurements. This is because such a threshold would change depending upon the noise in the observations. This provides a strong example for combining DI estimation with 3D modeling.

The 3D modeling was also applied to video sequences of a person performing different Yoga postures (Keogh, 2006). A representative image from five of the postures is shown in the top left corner of Fig. 4.

The deformable modeling algorithm requires the number of deformation modes (i.e., K) as the input. The estimated value of K using the proposed algorithm is shown in Fig. 4a–e for the five different postures.. The noise covariance was estimated using the approach described earlier on the gait dataset.[1]

We analyzed the estimation process for K against other approaches, and the reconstruction error for the 3D models. In Fig. 5a we plot the reprojection error for the 3D models against the number of basis shapes used (starting at two basis shapes). The reprojection error is the squared difference between the projection of the 3D model and the tracked features averaged over all the frames and feature points. We see that for each Yoga posture there is a cut-off value for the number of basis shapes, after which there is little or no decrease of the reconstruction error. However, the number of basis shapes needed to reach this cut-off value is significantly different ranging from four for

[1] There may be some wrong correspondences on the contour, though the amount of deviation in pixels is small. However, most of the correspondences in our examples are correct. The noise is calculated from the image gradients around each point and is not affected very much by a few pixels (due to the clean nature of this data and fixed background). Since the DI is derived from the most significant eigenvalues of the transformed observation matrix (Section 2.2), a few mismatches with a small deviation in pixels does not affect the estimate of the DI. Also, the DI will usually be applied in a learning phase and then used for further analysis. Thus correspondences can be chosen carefully so as to obtain an accurate estimate of the DI.

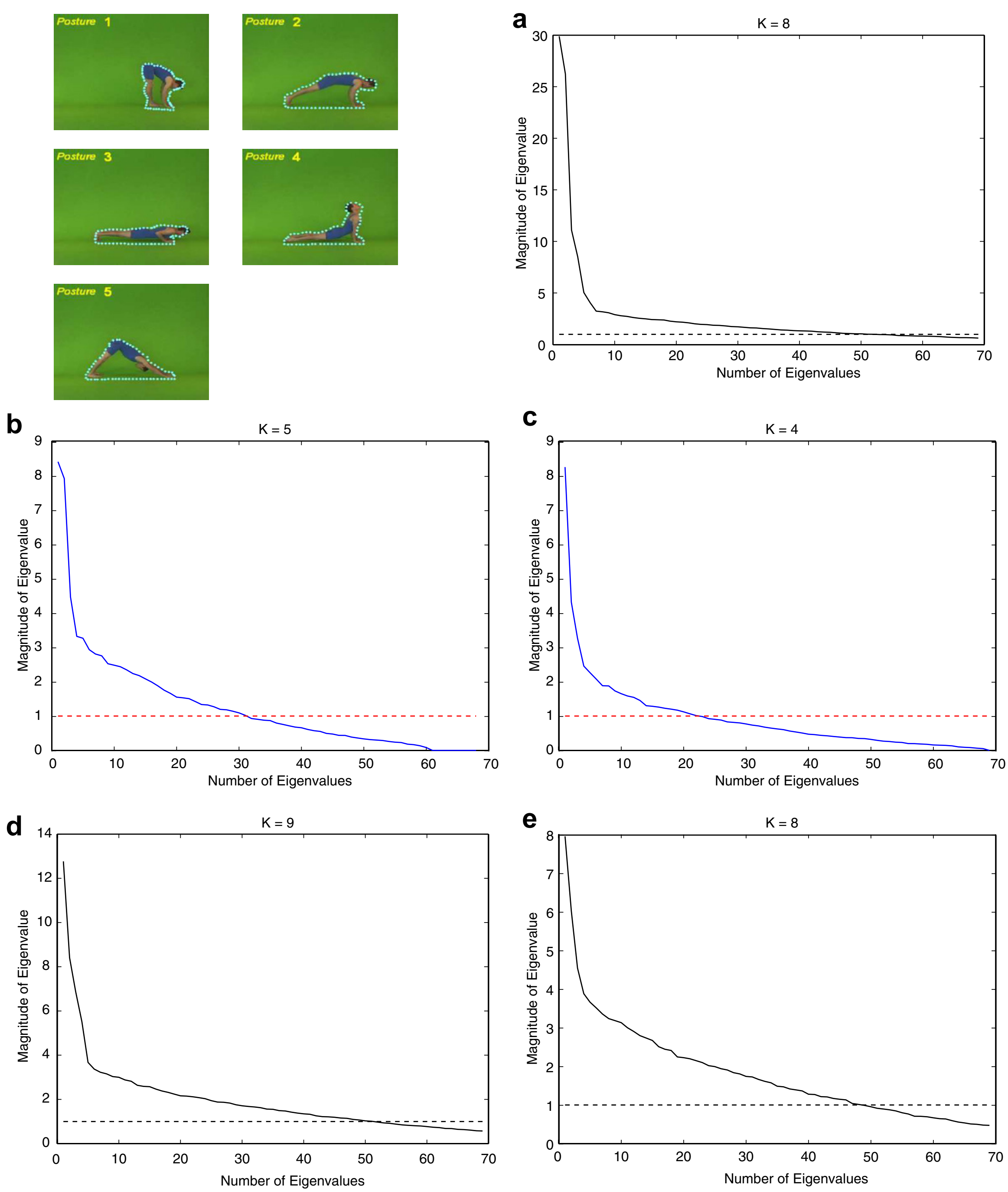

Fig. 4. One representative image for each posture with tracked features overlayed is shown on the top left. (a–e) Plot of the eigenvalues (magnitude vs. number of eigenvalues) of $\boldsymbol{R}_{\tilde{y}}$. The values of K, indicated on top of the plots, are 8, 5, 4, 9 and 8, respectively.

Posture 3 to nine for Posture 4. This is to be expected because the amount of deformation the body undergoes in each of these postures is very different. Comparing Fig. 5a with Fig. 4a–e, we see that the estimated number of basis shapes using our theory is close to this cut-off value. We consider this to be an "optimal" choice of the number of basis shapes since increasing it further increases the computational complexity without any significant gain in accuracy.

We now compare this result with existing methods for estimating the number of deformation modes. In most of these methods, this is done based on the rank of the measurement matrix, $\boldsymbol{W}$. We plot the eigenvalues of this matrix for each posture in Fig. 5b. We can see from the plots that choosing the "optimal" number of basis shapes requires different thresholds which are difficult to know a-priori. If we choose a fixed threshold, the numbers of basis shapes are nowhere near the optimal values. This underlines the

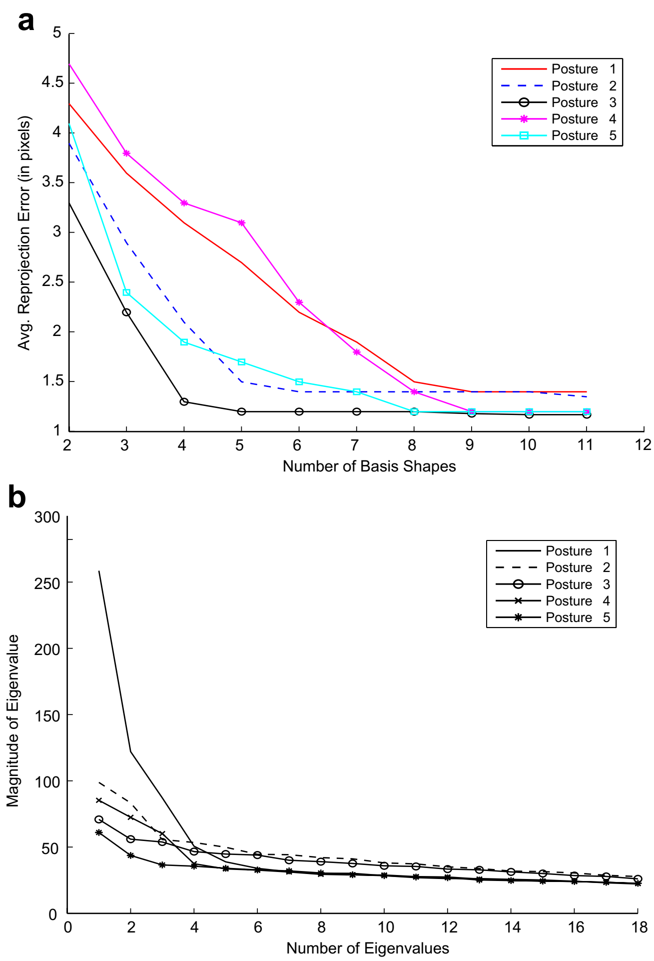

Fig. 5. (a) Plot of the distortion in the 3D reconstruction (measured in terms of the reprojection error averaged over all points and all frames) vs. the number of basis shapes used for the 3D model estimation. (b) Plot of the eigenvalues of $\boldsymbol{R}_y$, i.e., the original measurement matrix.

difficulty of choosing the number of deformation modes based on the measurement matrix.

We want to note here that there are many accurate methods for reconstructing 3D models of articulated objects. Some recent examples are (Tresadern and Reid, 2005; Yan and Pollefeys, 2005). We do not propose alternative methods for 3D modeling. However, the rank constraints that are derived from the deformable contour analysis are retained in the more complex modeling methods and can be used for reconstructing articulated bodies.

4. Conclusions

In this paper, we have a presented a method for estimating the deformability of shape sequences obtained from a sequence of video frames. We assumed an affine camera projection model and that a deformable shape can be represented using a linear combination of basis shapes. The theory relied on estimating the number of 3D basis shapes from the 2D feature positions representing the shapes. The deformability index is directly related to the number of basis shapes. The computation of the deformability index can handle rigid 3D transformations of the shape, though it does not require prior estimation of the 3D structure or motion. Our method takes into account the statistics of the noise in the feature positions. We presented experimental results in human movement analysis using motion capture and real-life video images. The estimates of the deformability index are in accordance with what would be expected intuitively by a human observer and corroborate certain hypotheses in the existing literature on human motion analysis. Results on 3D deformable object modeling and comparison with existing approaches are also presented.

Acknowledgement

The author would like to thank Rama Chellappa for many interesting discussions related to this work.

References

Brand, M., 2001. Morphable 3d models from video. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition.

Brand, M., 2005. A direct method for 3d factorization of nonrigid motion observed in 2d. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition.

Daniilidis, K., Spetsakis, M. 1993. Understanding noise sensitivity in structure from motion. In: VisNav93.

Del Bue, A., Llad, X., Agapito, L., 2005. Non-rigid face modelling using shape priors. In: Analysis and Modelling of Faces and Gestures, Lecture Notes in Computer Science.

Irani, M., 1999. Multi-frame optical flow estimation using subspace constraints. In: Proc. IEEE Internat. Conf. on Computer Vision.

Kale, A., Rajagopalan, A., Sundaresan, A., Cuntoor, N., Roy-Chowdhury, A., Krueger, A., Chellappa, R., 2004. Identification of humans using Gait. IEEE Trans. Image Process., 1163–1173.

Kanatani, K., 1996. Statistical Optimization for Geometric Computation: Theory and Practice. North-Holland.

Kanatani, K., 2004. Uncertainty modeling and model selection for geometric inference. IEEE Trans. Pattern Anal. Machine Intell. 26, 1307–1319.

Keogh, E., 2006. The UCR time series data mining archive. <http://www.cs.ucr.edu/~eamonn/TSDMA>.

Lee, L., Grimson, W., 2002. Gait analysis for recognition and classification. In: Proc. IEEE Conf. on Face and Gesture Recognition.

Liu, C., Ahuja, N., 2004. A model for dynamic shape and its applications. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition.

Roy-Chowdhury, A., 2005. A measure of deformability of shapes, with applications in human motion analysis. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition.

Sarkar, S., Phillips, P., Liu, Z., Vega, I., Grother, P., Bowyer, K., 2005. The HumanID Gait challenge problem: Data sets, performance, and analysis. IEEE Trans. Pattern Anal. Machine Intell. 27 (2), 162–177.

Soatto, S., Yezzi, A., 2002. DEFORMOTION: Deforming motion, shape average and the joint registration and segmentation of images. In: Proc. European Conf. on Computer Vision.

Stoica, P., Moses, R., 1997. Introduction to Spectral Analysis. Prentice-Hall.

Sun, Z., Ramesh, V., Tekalp, A., 2001. Error characterization of the factorization method. Computer Vision and Image Understanding 82, 110–137.

Tanawongsuwan, R., Bobick, A., 2002. Modelling the effects of walking speed on appearance-based gait recognition. In: Computer Vision and Pattern Recognition.

Tomasi, C., Kanade, T., 1992. Shape and motion from image streams under orthography: A factorization method. Int. J. Comput. Vision 9, 137–154.

Torresani, L., Yang, D., Alexander, E., Bregler, C., 2001. Tracking and modeling non-rigid objects with rank constraints. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition.

Torresani, L., Hertzmann, A., Bregler, C., 2003. Learning non-rigid 3d shape from 2d motion. In: Neural Information Processing Systems.

Tresadern, P., Reid, I., 2005. Articulated structure from motion by factorization. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition.

Trucco, A., Verri, A., 1998. Introductory Techniques for 3D Computer Vision. Prentice-Hall.

Vaswani, N., Roy-Chowdhury, A., Chellappa, R., 2003. Activity recognition using the dynamics of the configuration of interacting objects. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition.

Veeraraghavan, A., Roy-Chowdhury, A., Chellappa, R., 2005. Matching shape sequences in video with applications in human movement analysis. IEEE Trans. Pattern Anal. Machine Intell., 1896–1909.

Weng, J., Huang, T., Ahuja, N., 1987. 3-D motion estimation, understanding, and prediction from noisy image sequences. IEEE Trans. Pattern Anal. Machine Intell. 9, 370–389.

Xiao, J., Chai, J., Kanade, T., 2004. A closed-form solution to non-rigid shape and motion recovery. In: Proc. European Conf. on Computer Vision.

Yan, J., Pollefeys, M., 2005. A factorization-based approach to articulated motion recovery. In: Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition.