Recognizing Walking People

Stefan Carlsson

Numerical Analysis and Computing Science, Royal Institute of Technology, (KTH),
S-100 44 Stockholm, Sweden, stefanc@nada.kth.se,
http://www.nada.kth.se/~stefanc

Abstract. We present a method for recognition of walking people in monocular image sequences based on extraction of coordinates of specific point locations on the body. The method works by comparison of sequences of recorded coordinates with a library of sequences from different individuals. The comparison is based on the evaluation of view invariant and calibration independent *view consistency constraints*. These constraints are functions of corresponding image coordinates in two views and are satisfied whenever the two views are projected from the same 3D object. By evaluating the view consistency constraints for each pair of frames in a sequence of a walking person and a stored sequence we get a matrix of consistency values that ideally are zero whenever the pair of images depict the same 3D-posture. The method is virtually parameter free and computes a consistency residual between a pair of sequences that can be used as a distance for clustering and classification. Using interactively extracted data we present experimental results that are superior to those of previously published algorithms both in terms of performance and generality.

Keywords: structure from motion, calibration, object recognition

1 Introduction

Visual analysis of human motion is an area with a large potential for applications. These include medical analysis, automatic user interfaces, content based video analysis etc. It has therefore received an increased amount of attention during recent years [1], [10]. A dominant part of the work in the field has been concerned with automatic tracking using shape models in 2D or 3D. [2], [5], [11], [14], [15], [17] with the main application of classifying human action in mind. For this, a 3D model would be invaluable which in general requires multiple views for the analysis. The important case of using just a single view makes the problem of 3D model acquisition far more difficult and has so far met with only limited success [10] This difficulty of automatic 3D analysis from a single sequence stands in sharp contrast to the ease at which we can perceptually interpret human motion and action, even from very impoverished monocular stimuli as moving light displays [12] although the ability to recognize a specific individual from such displays is far from perfect [9].

D. Vernon (Ed.): ECCV 2000, LNCS 1842, pp. 472–486, 2000.

The purpose of the work presented in this paper is the identification of walking people from a single monocular video sequence acquired from an arbitrary relative viewpoint. In order for the identification to be insensitive to changes in viewpoint and in order to overcome the perspective effects that will be present during a walking cycle, we should ideally make use of 3D properties of human gait such as time sequences of relative angles of limbs etc. which as was pointed out is a very difficult problem. All work so far in automatic person identification from image sequences have therefore restricted the motion to frontoparallel walking relative to the camera. [13], [14].

Since 3D model acquisition from a single view is notoriously difficult and unstable it will require strong model assumptions about the movement in 3D in order to get a stable solution. Making strong assumptions about the movement in 3D is however highly undesirable if the application is identification since the model assumptions will wipe out the individual differences between the different persons that are to be identified.

In opposition to this we will allow for very general relative viewpoint by basing the recognition implicitly on the 3D structure and motion of the walking person without performing any explicit 3D reconstruction. This will be done by exploiting *view consistency constraints* that exists between two views of the same 3D structure. Given two image point sets, the view consistency constraints are functions of the coordinates of the points in the two images and they are satisfied if the two images are the projection of the same 3D point set. They therefore answer the question: "can these two views be the projection of the same 3D structure ?"

In the following sections we will present the geometric interpretation and algebraic derivation of VC-constraints and demonstrate how they can be used in order to verify whether two image sequences depict the same walking person.

2 View Consistency Constraints

2.1 Geometric Interpretation

Whenever we have two images of the same point set in 3D, the two image planes can be positioned so that the lines of sights through the points intersect at the coordinates of the 3D point set (fig. 1 A) We will refer to two such image point sets as being *view consistent*.

Note that view consistency follows from the fact that two views are projected from the same 3D point set but the reverse is not true. Having two view consistent point sets does not imply that they are projected from the same 3D point set. Two different 3D point sets can align accidentally to produce two view consistent image point sets. The equivalence class of ambiguous 3D point sets with this property can be generated easily by just extending the lines of sights in the two images of fig 1 B. If however we are viewing a restricted class of 3D shapes

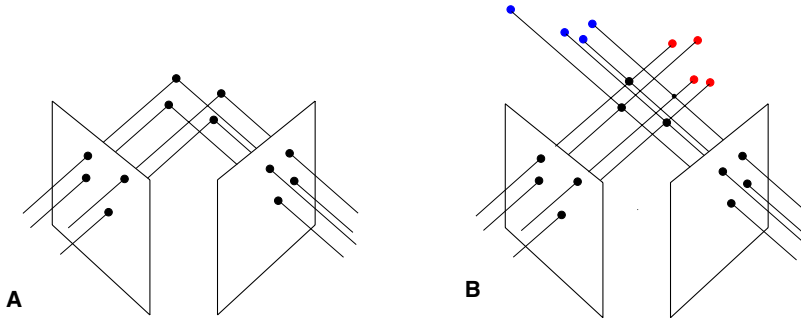

Fig. 1. A). Two image point sets are view consistent if they are the projection of the same 3D point set. B).Two image point sets can be accidentally view consistent although they are not projected from the same 3D structure

with statistically a priori defined properties, it is likely that these ambiguous equivalent classes of 3D point sets can be narrowed down substantially. If the points represent specific locations on the human body e.g. the constraints on human posture will lead to a restriction of the 3D shape ambiguity and the degree to which the view consistency constraint is satisfied will reflect the difference in 3D shape of the point sets giving rise to the two views. The degree to which this is true must of course eventually be determined by experimental evaluation for each restricted class of shapes.

2.2 Two View Consistency

The fundamental constraints relating the projection of a 3D point to an image involves three kinds of parameters: Camera parameters, 3D coordinates and image coordinates. These are related by a projection equation, which varies with the kind of geometry and degree of calibration assumed for the camera. If more than one view is available, the 3D coordinates can be eliminated, leaving constraint relations between camera parameters and image coordinates in the two views. These are known as epipolar constraints. Alternatively, by having sufficiently many points in one view, the camera parameters can be eliminated leaving constraints in 3D and image coordinates known as single view shape constraints. [6], [8], [16], [18], The process of elimination can be continued from these constraints. By using sufficiently many points the camera parameters in the epipolar constraints can be eliminated leaving just constraints in the image coordinates of the two views. Alternatively by using multiple views, the 3D coordinates can be eliminated from the single view shape constraints, leaving identical constraints in the image coordinates. These image coordinate constraints reflect the fact that the two views are the projection of the same 3D point set and are therefore the algebraic expressions of the view consistency constraints. [1]

[1] Since the initial submission of this paper it has come to the author's attention that the specific four point constraint for the case of known scale factors was actually first derived in [3] and discussed for use in recognition applications in [4]

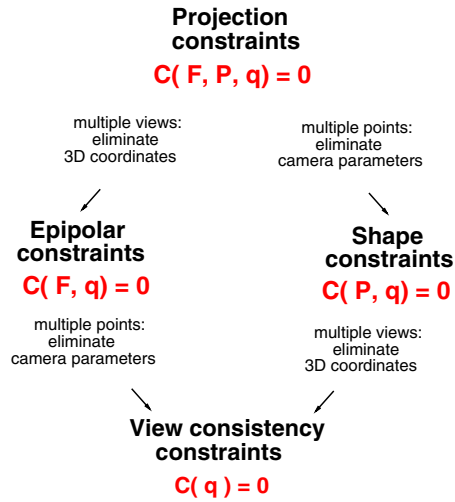

Fig. 2. Derivation of view consistency constraints by elimination involving camera parameters: **F**, 3D coordinates: **P** and image coordinates: **q**

In app. 1 it is shown that for orthographic projection cameras with unknown scale factors, four corresponding points with image coordinates p_i^a and p_i^b in the two views respectively satisfy the polynomial consistency constraint

$$\begin{aligned} &\alpha \ \left([\ B_2\ A_3\ A_4\ ] + [\ A_2\ B_3\ A_4\ ] + [\ A_2\ A_3\ B_4\ ] \right) - \\ -\ &\beta \ \left([\ B_2\ B_3\ A_4\ ] + [\ B_2\ A_3\ B_4\ ] + [\ A_2\ B_3\ B_4\ ] \right) \ = \ 0 \end{aligned} \tag{1}$$

where the determinants $[\ A_i\ A_j\ B_k\ ]$ are defined as:

$$[\ A_i\ A_j\ B_k\ ] \ = \ \begin{bmatrix} < a_i\ a_2 > & < a_j\ a_2 > & < b_k\ b_2 > \\ < a_i\ a_3 > & < a_j\ a_3 > & < b_k\ b_3 > \\ < a_i\ a_4 > & < a_j\ a_4 > & < b_k\ b_4 > \end{bmatrix} \tag{2}$$

with $a_i = p_i^a - p_1^a$ and $b_i = p_i^b - p_1^b$ $i = 2 \ldots 4$ and α and β are the unknown scale factors squared of the two cameras. By taking a fifth point we can eliminate the ratio α/β to get a view consistency constraint polynomial in five points, as shown in app. 1.

In general we will have more than five points available however. It is therefore more effective to eliminate α and β using a regression procedure. based on four point constraint polynomials of type 1. These polynomials will not all be independent. For our application we interactively select 8 points of the image of a walking person, fig 3.

All combinations of four points give a polynomial constraint acc. to eq 1:

$$\alpha\ P_1(i,j,k,l) - \beta\ P_2(i,j,k,l) \ = \ 0 \tag{3}$$

It can be shown theoretically and was verified experimentally that choosing point sets with 3 points collinear or close to, will result in numerical instability.

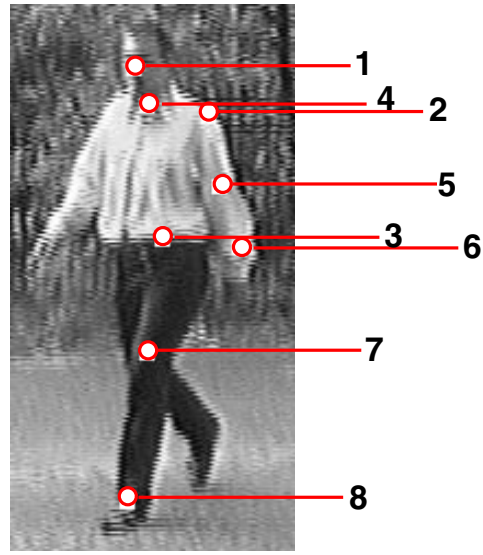

Fig. 3. Interactively selected point locations on human body

In order to avoid as far as possible having triplets of 3 collinear points we choose points 1 2 3 in all combinations and vary the fourth points among the set 4, 5, 6, 7, 8. We then get the set of constraint polynomials:

$$\alpha\ P_1(1,2,3,i) - \beta\ P_2(1,2,3,i)\ =\ 0 \qquad\qquad i = 4,5,6,7,8 \tag{4}$$

The values of α and β are found by regression:

$$\min_{\alpha^2+\beta^2=1} \sum_{i=4}^{8} (\alpha\ P_1(1,2,3,i) - \beta\ P_2(1,2,3,i)\ )^2 \tag{5}$$

and the regression residual using optimal values of α and β

$$R(\hat{\alpha},\hat{\beta})\ =\ \sum_{i=4}^{8} (\hat{\alpha}\ P_1(1,2,3,i) - \hat{\beta}\ P_2(1,2,3,i)\ )^2 \tag{6}$$

is used as a measure of the degree of consistency of the two views. This value is zero if the two views are noise free orthographic projections of the same 3D point set. If the views are projections of different point sets, this value will in general deviate from zero. In general we can expect this value to measure the degree of 3D similarity of two point sets. Especially when we are considering a restricted class of 3D shapes given by human posture.

2.3 Sequence Consistency

The two view consistency constraint residual is the basis for the algorithm for recognition of walking people. For this we will use sequences of one walking

cycle from different individuals. With normal walking speed this means about 27 - 29 frames using 25 Hz frame rate. By using all frames in the sequence we will get a statistically more robust identification compared to just two views. The procedure for computing a measure of consistency of two sequences is as follows:

1. The image coordinates of 8 specific body locations are extracted interactively from each frame in the two sequences a and b.

2. For each pair of frames , one from sequence a and the other from sequence b we compute the regression residual $R(\hat{\alpha}, \hat{\beta})$ of eq 6. If the sequences depict the same person, we will get low values for every pair of frame with the same 3D-posture. Depending on the synchronization and relative walking speed in the two sequences low values will show up along a parallel displaced and tilted diagonal line in the view consistency matrix composed of all the pairwise regression residuals.

3. Regression residuals are averaged along lines in the view consistency matrix with various starting and stopping pairs marked in fig 4. The minimum average value will be referred to as the *sequence consistency residual.*

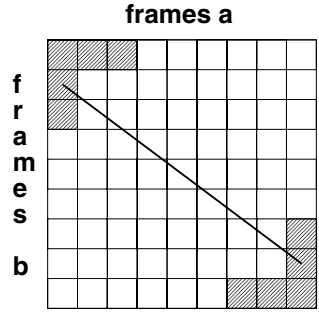

Fig. 4. Depending on syncronisation and relative walking speed in two sequences a and b, minimum sequence consistency residual values will appear along a tilted diagonal line in the VCR-matrix. Minimum average value of the VCR's is sought for among the lines with start and stop positions in the marked areas.

Fig 5 shows examples of 16 view consistency matrices for four sequences compared with each other. The first two Ka-1 and Kb-1 depict person K from different viewpoints and the second two depict person M, also from different viewpoints. The first and the last frames of the sequences can be found in fig 10. The sampling of the frames in the sequences was chosen in order to get approximate synchrony of the walking cycles although this is not critical but helps the visualization of the view consistency matrices. From fig 5 we note that

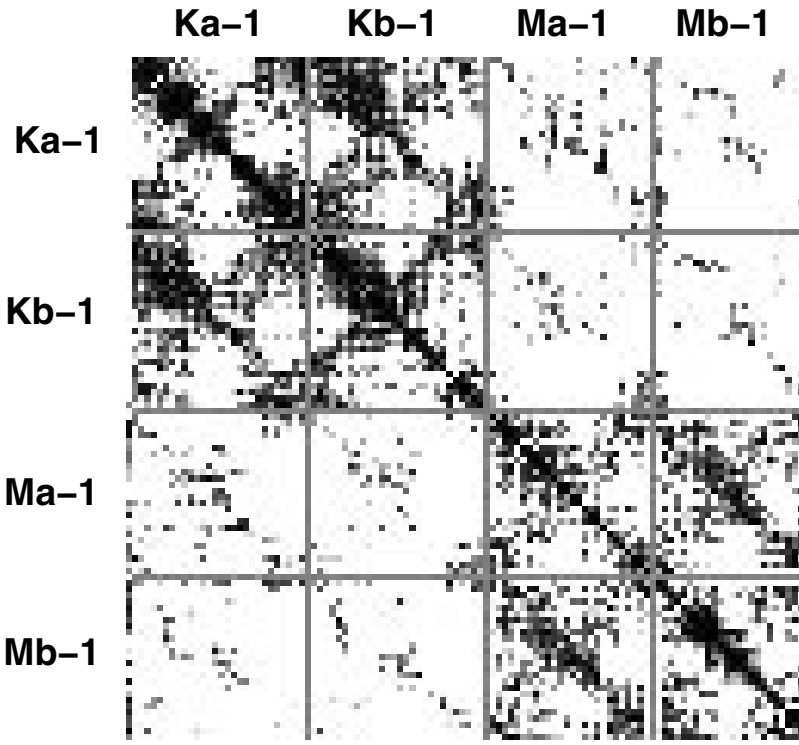

Fig. 5. View consistency residual (VCR)-matrices for sequences Ka Kb Ma Mb compared with each other. Dark = low residuals, light = high residuals

we get low residual values along and close to the diagonal of the concatenated view consistency matrices. The values along the diagonal is of course zero since exactly the same sequence with exactly the same frames are compared there. The width of the low values close to the diagonal varies depending on the speed of motion during the walking cycle. The narrowest parts can be seen in the phase of the motion of the left leg which gives the fastest variation of the 3D posture. We see that we also get low residual values of the matrices depicting the same person but different walking sequences (Ka-1 Kb-1) and (Ma-1 Mb-1) but substantially higher residual values for the matrices where different persons are compared.

3 Experimental Results and Conclusions

3.1 Recording of Walking Sequences

In order to test whether the sequence consistency measure defined in the previous chapter can be used for identification , we recorded sequences of walking people with varying direction of motion relative to the camera line of sight (fig. 6.)

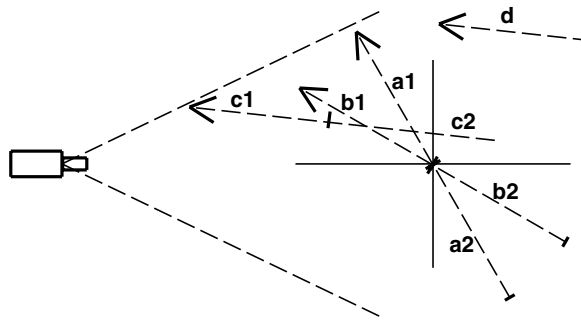

Fig. 6. Recorded segments of one walking cycle in various relative directions

The attempt was made to record walking directions with angles 30, and 60 degrees relative to the image plane. They are denoted as a and b respectively in fig. 6. Sequences c and d are about 80 degrees relative walking direction. Every segment in fig. 6 represents one walking cycle. Two consecutive walking cycles were extracted when possible. These are denoted as a1, a2 , b1 b2 and c1, c2 respectively. Six different persons , denoted A, T, C, D, K and M were recorded. Sequences of one walking cycle were extracted from the recorded material with an attempt to choose temporal syncronisation for purposes of display. Fig 10 shows the first and last frames of the sequences a1 and b1 for four different persons. The choice of syncronisation of the sequences together with the fact that walking directions were somewhat approximate relative to the attempted means that there is a slight deviation from the ideal geometry of fig. 6. The two persons A and T were recorded at another occasion in the directions c and d.

For each recorded sequence, 8 points according to fig. 3 were selected interactively. The view angle of the camera was around 45 deg. and fig 10 shows the actual frames recorded.

3.2 Sequence Consistency Residuals

For the approximate directions of 30 and 60 degrees relative to the image plane, denoted a and b respectively, we were able to extract 14 different sequences of one walking cycle from persons C, D, K and M and for directions c and d, 6 sequences from persons A and T. They were all compared pairwise and the sequence consistency residuals acc. to section 2.2 was computed. These are shown in the table of fig 7

The table shows that sequence consistency residuals are in general substantially lower for sequence pairs of the same person compared to sequence pairs of different persons. The sequences of a certain person can therefore be visualized as clusters by the method known as multidimensional scaling where we try to plot all sequences as points in a 3D-space in order that the interpoint distances should equal the sequence consistency residuals. This is in general not possible to do consistently in an exact way. Fig 8 shows a plot of this for all walking segments of all the six persons in the experiments which is *conservative* in the

	A-c1	A-c2	A-d	T-c1	T-c2	T-d	C-a1	C-a2	C-b1	D-a1	D-a2	D-b1	D-b2	K-a1	K-a2	K-b1	K-b2	M-a1	M-a2	M-b1
A-c1	0.0	2.6	2.2	5.1	8.4	6.2	34.4	27.8	14.6	13.1	11.2	8.9	7.9	24.4	12.8	10.2	10.6	15.5	22.0	9.3
A-c2	2.6	0.0	3.3	10.4	17.9	7.9	41.0	38.4	23.2	11.0	8.5	6.4	6.3	25.3	17.9	12.0	10.6	8.9	10.9	6.9
A-d	2.3	3.3	0.0	6.9	10.7	3.0	17.3	14.5	8.7	6.0	5.9	4.8	5.4	11.4	6.4	4.4	5.3	7.8	12.2	7.1
T-c1	4.9	10.5	6.8	0.0	2.1	2.1	11.6	10.3	4.3	21.6	17.1	8.6	11.7	24.7	12.3	7.4	13.0	32.9	38.8	26.4
T-c2	6.8	17.6	10.0	2.2	0.0	3.0	5.3	9.2	4.2	17.6	20.6	7.9	8.1	17.2	9.6	9.3	11.9	24.2	30.0	20.4
T-d	6.0	7.9	3.0	2.5	3.1	0.0	11.5	8.6	4.5	17.5	16.9	9.9	7.7	19.3	8.1	6.3	6.8	31.0	36.7	18.4
C-a1	33.6	41.2	18.8	12.6	5.4	13.2	0.0	7.2	6.9	24.6	15.2	17.1	17.0	16.5	15.6	9.2	19.4	33.6	34.2	42.0
C-a2	23.7	38.7	14.5	10.3	9.2	8.6	7.2	0.0	7.3	32.5	22.5	12.0	13.2	10.1	6.3	6.8	12.1	33.4	24.1	27.9
C-b1	14.6	23.5	8.9	4.3	4.2	4.6	6.7	7.3	0.0	35.4	33.4	15.7	14.3	27.4	13.2	13.1	16.3	42.9	24.0	44.3
D-a1	13.8	11.8	6.6	21.7	17.9	18.7	24.6	32.8	35.4	0.0	1.9	5.4	5.4	21.9	20.2	16.5	15.4	20.5	12.6	15.7
D-a2	11.7	9.2	6.6	22.8	21.5	20.3	15.2	22.6	34.2	1.9	0.0	3.5	2.7	16.6	14.8	12.5	12.7	13.9	8.7	10.1
D-b1	9.5	6.4	4.8	11.6	9.7	10.9	17.1	12.3	16.3	5.4	3.2	0.0	2.9	16.4	12.3	9.2	8.7	12.5	8.3	9.1
D-b2	8.2	6.7	6.3	11.8	8.2	8.2	17.1	13.2	14.3	5.2	2.7	2.9	0.0	20.0	9.3	9.4	6.1	10.9	4.3	7.8
K-a1	24.4	23.1	11.4	27.0	18.1	21.8	16.5	10.6	28.3	22.2	16.6	16.1	20.6	0.0	5.8	4.9	9.2	36.3	37.9	20.8
K-a2	12.6	16.6	6.4	12.2	9.6	9.0	15.3	6.3	13.2	19.8	14.5	12.2	9.4	5.8	0.0	4.1	4.8	17.3	19.2	16.3
K-b1	10.1	12.0	4.4	9.2	9.5	6.8	9.2	6.8	13.2	16.6	12.5	9.2	9.4	4.9	4.2	0.0	4.7	23.8	18.6	12.0
K-b2	10.4	10.1	5.2	14.6	13.1	8.7	19.4	12.0	16.8	15.0	12.7	8.3	6.1	9.2	4.7	4.5	0.0	17.1	11.3	12.5
M-a1	22.9	14.2	13.3	49.5	32.0	40.4	34.1	33.6	43.6	20.1	14.6	13.0	10.9	36.3	21.7	23.8	17.2	0.0	3.7	5.7
M-a2	11.4	7.7	7.9	27.3	20.4	19.1	42.0	24.0	24.1	12.7	8.7	8.3	4.3	37.9	19.2	18.6	11.3	3.3	0.0	4.6
M-b1	19.4	10.3	8.9	42.4	25.9	36.2	33.7	28.8	45.3	15.7	10.0	9.8	8.2	20.6	16.3	11.9	13.0	5.6	4.6	0.0

Fig. 7. Sequence consistency residuals for all sequence pairs

sense that the lowest sequence consistency residuals, those < 10, e.g. all residuals corresponding to the same person are reproduced more or less exactly while the large residuals are in general under estimated in the plot.

From this plot we see that walking sequences from the same person in various different walking directions can be grouped into non-overlapping clusters. For reasons of visualization, the dimension of the plot is chosen as 3. We do not really know the actual intrinsic dimensionality of the data set, so the conservative 3 capr-D plot gives in a sense the worst case projection of the data from some unknown high dimensional space onto a 3 dimensional space.

3.3 Conclusions

The performance of any algorithm for classification of data into separate classes depends of course on the number of classes, in our case the number of different individuals. It also depends on how many sequences that are stored as reference for each individual. Given a set of training sequences in a library, one can ask for the probability of a new recorded sequence to be classified correctly. We can test this on our material using a so called "leave one out" procedure, whereby one sequence at a time is considered as the sequence to be classified and the remaining sequences are all considered to be reference sequences. We can then use a nearest neighbor classifier based on computing the average distance to all reference sequences from a certain person. If this is done we get the result of the table in fig 9.

From this table we see that of 20 sequences, 19 are classified correctly using a simple minimum distance classifier. The only exception is sequence C-b1 being classified as T. This is of course a promising result but should be seen in the

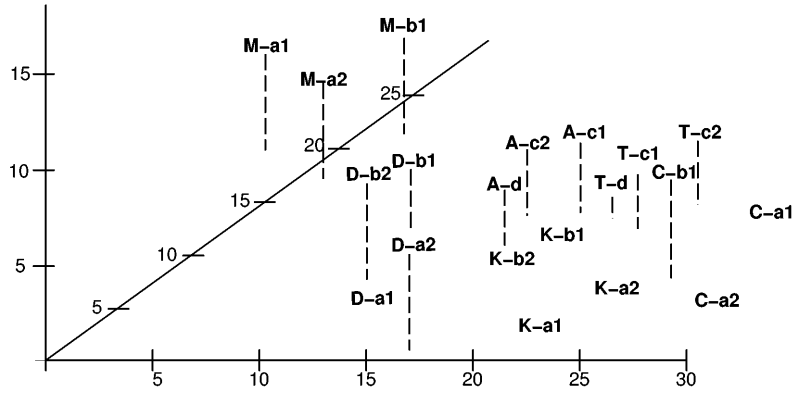

Fig. 8. 3D multidimensional scaling plot of sequence consistency residuals. The interpoint distances in 3D correspond roughly to the sequence consistency residuals of fig 7.

light of the fact that we only have 20 sequences from 6 persons. Increasing the number of persons would of course increase the chance of misclassification.

It is interesting to note in comparison that similar procedures for computing recognition rates led to 81 % in [14] for 26 sequences of 5 persons and around 90 % in [13] for 42 sequences of 6 persons. Both these cases used automatic data extraction but were restricted to frontoparallel walking however while our algorithm , using interactive feature extraction, performs over a wide range of relative walking directions. Using view consistency constraints for identification of walking people, we therefore can claim superior results and more generality at the price of using interactive selection of feature points.

	A-c1	A-c2	A-d	T-c1	T-c2	T-d	C-a1	C-a2	C-b1	D-a1	D-a2	D-b1	D-b2	K-a1	K-a2	K-b1	K-b2	M-a1	M-a2	M-b1
A	2.4	3.0	2.8	7.4	11.5	5.6	31.2	25.6	15.7	10.7	9.2	6.9	7.1	19.6	11.9	8.8	8.6	16.8	9.0	12.9
T	6.6	12.1	6.9	2.1	2.6	2.8	10.4	9.4	4.4	19.4	21.5	10.7	9.4	22.3	10.3	8.5	12.1	40.6	22.3	34.8
C	25.6	34.2	13.5	8.7	6.2	8.2	7.1	7.2	7.0	30.9	24.0	15.2	14.9	18.5	11.6	9.7	16.1	37.1	30.0	35.9
D	10.3	8.0	5.5	4.8	13.6	13.0	18.5	20.1	24.7	4.2	2.7	3.8	3.6	18.9	14.0	11.9	10.5	14.7	8.5	10.9
K	14.5	16.4	6.9	4.3	12.0	10.1	15.2	8.8	17.5	18.5	14.2	11.6	11.2	6.6	4.9	4.6	6.1	24.8	21.7	15.5
M	15.6	8.9	9.0	2.7	24.9	28.7	36.6	28.5	37.1	16.3	10.9	10.0	7.7	31.7	17.6	18.1	13.6	4.7	3.9	5.1

Fig. 9. Average sequence consistency residuals over different individuals for all sequences

In its present form the we believe that the algorithm can be useful in e.g. forensic science applications where the problem often is to classify a single sequence and the library is limited.

Acknowledgement
This work was supported by the EU Esprit Project 23515 IMPROOFS.

4 Appendix

4.1 View Consistency Constraints - Scaled Orthographic Projection

For most real world cameras we can assume to have orthographic projection and square pixels, i.e. the same scale factor in x and y. This only unknown internal camera parameter is then a scale factor σ. The projection equation for an arbitrary orthogonal image coordinate system for this camera can be written as:

$$\begin{aligned} x &= \sigma\, r_1^T \bar{P} + x_0 \\ y &= \sigma\, r_2^T \bar{P} + y_0 \end{aligned} \tag{7}$$

where r_1 and r_2 are the first two rows of an arbitrary 3×3 rotation matrix. We will now derive the view consistency constraints for two cameras of this kind by elimination of all camera and 3D shape parameters. Note that two having views does not permit the explicit determination of relative camera rotation but this will be of no concern for the view consistency constraints since rotation is to be eliminated anyway.

The unit vectors of the orthonormal rotation matrix r_1, r_2 and r_3 can be used to expand the vector $\bar{P}_i - \bar{P}_1$ Introducing the unknown parameter $\gamma_i = r_3^T(\bar{P}_i - \bar{P}_1)$ and taking differences to eliminate the constants x_0, y_0, we get:

$$\begin{aligned} \sigma^{-1}(x_i - x_1) &= r_1^T(\bar{P}_i - \bar{P}_1) \\ \sigma^{-1}(y_i - y_1) &= r_2^T(\bar{P}_i - \bar{P}_1) \\ \gamma_i &= r_3^T(\bar{P}_i - \bar{P}_1) \end{aligned} \tag{8}$$

Using the orthonormality of the matrix $(r_1\ r_2\ r_3)$ we get:

$$\bar{P}_i - \bar{P}_1 = \sigma^{-1}(x_i - x_1)\, r_1 + \sigma^{-1}(y_i - y_1)\, r_2 + \gamma_i\, r_3 \tag{9}$$

By taking inner products we can eliminate the rotation matrix:

$$\begin{aligned} &< (\bar{P}_i - \bar{P}_1)\ (\bar{P}_j - \bar{P}_1) > = \\ &= \sigma^{-2}(x_i - x_1)\ (x_j - x_1) + + \sigma^{-2}(y_i - y_1)\ (y_j - y_1) + \gamma_i\, \gamma_j \end{aligned} \tag{10}$$

Using image data from two views a and b we can eliminate the 3D coordinates and get:

$$\sigma^{a-2}\,(x_i^a\;-\;x_1^a)\,(x_j^a\;-\;x_1^a)\;+\;\sigma^{a-2}\,(y_i^a\;-\;y_1^a)\,(y_j^a\;-\;y_1^a)\;+\;\gamma_i^a\,\gamma_j^a\;=$$

$$\sigma^{b-2}\,(x_i^b\;-\;x_1^b)\,(x_j^b\;-\;x_1^b)\;+\;\sigma^{b-2}\,(y_i^b\;-\;y_1^b)\,(y_j^b\;-\;y_1^b)\;+\;\gamma_i^b\,\gamma_j^b$$

which we write as:

$$\begin{aligned}\rho(\;(x_i^b\;-\;x_1^b)\,(x_j^b\;-\;x_1^b)\;+\;(y_i^b\;-\;y_1^b)\,(y_j^b\;-\;y_1^b)\;)\;-\\ -\;(x_i^a\;-\;x_1^a)\,(x_j^a\;-\;x_1^a)\;+\;(y_i^a\;-\;y_1^a)\,(y_j^a\;-\;y_1^a)\;=\\ =\;\sigma^{a2}(\;\gamma_i^a\,\gamma_j^a\;-\;\gamma_i^b\,\gamma_j^b\;)\end{aligned}\tag{11}$$

where we have used $\rho = (\sigma_a/\sigma_b)^2$

In order to get a more compact notation we write the inner products as:

$$\begin{aligned}<a_i\;a_j>\;&=\;(x_i^a\;-\;x_1^a)\,(x_j^a\;-\;x_1^a)\;+\;(y_i^a\;-\;y_1^a)\,(y_j^a\;-\;y_1^a)\\ <b_i\;b_j>\;&=\;(x_i^b\;-\;x_1^b)\,(x_j^b\;-\;x_1^b)\;+\;(y_i^b\;-\;y_1^b)\,(y_j^b\;-\;y_1^b)\;)\end{aligned}\tag{12}$$

$$\alpha_i\;=\sigma^a\gamma_i^a\qquad\beta_i\;=\sigma^a\gamma_i^b\tag{13}$$

Using this we get: Using this we get:

$$\rho<b_i\;b_j>\;-\;<a_i\;a_j>=\;\alpha_i\alpha_j\;-\beta_i\beta_j\tag{14}$$

where i and j ranges over $2\ldots n$ where n is the number of points. By using sufficiently many points we will now eliminate the unknowns, σ, α_i and β_i to get view consistency constraints expressed in terms of the image coordinates$<$ $a_i\;a_j>,<b_i\;b_j>$ of the two views only.

Consider first the case of having the same scale factor in both views, $\sigma^a = \sigma^b \Longrightarrow \rho = 1$ and take four points. We then have to eliminate six unknown parameters $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3$. In general we can write:

$$\begin{pmatrix}\alpha_4\\ \\ \beta_4\end{pmatrix}\;=q_1\;\begin{pmatrix}\alpha_2\\ \\ \beta_2\end{pmatrix}\;+\;q_2\;\begin{pmatrix}\alpha_3\\ \\ \beta_3\end{pmatrix}\tag{15}$$

Taking inner products of both sides of this equation with vectors $(\alpha_2, -\beta_2)$, $(\alpha_3, -\beta_3)$, and $(\alpha_4, -\beta_4)$ we get the linear system of equations:

$$\begin{aligned}\alpha_4\alpha_2\;-\beta_4\beta_2\;&=\;q_1\;(\alpha_2\alpha_2\;-\beta_2\beta_2)\;+\;q_2\;(\alpha_3\alpha_2\;-\beta_3\beta_2)\\ \alpha_4\alpha_3\;-\beta_4\beta_3\;&=\;q_1\;(\alpha_2\alpha_3\;-\beta_2\beta_3)\;+\;q_2\;(\alpha_3\alpha_3\;-\beta_3\beta_3)\\ \alpha_4\alpha_4\;-\beta_4\beta_4\;&=\;q_1\;(\alpha_2\alpha_4\;-\beta_2\beta_4)\;+\;q_2\;(\alpha_3\alpha_4\;-\beta_3\beta_4)\end{aligned}\tag{16}$$

Substituting the α and β expressions using eq. 14 we get:

$$\begin{aligned}
&< a_4a_2 > - < b_4b_2 > = q_1(< a_2a_2 > - < b_2b_2 >) + q_2(< a_3a_2 > - < b_3b_2 >) \\
&< a_4a_3 > - < b_4b_3 > = q_1(< a_2a_3 > - < b_2b_3 >) + q_2(< a_3a_3 > - < b_3b_3 >) \\
&< a_4a_4 > - < b_4b_4 > = q_1(< a_2a_4 > - < b_2b_4 >) + q_2(< a_3a_4 > - < b_3b_4 >)
\end{aligned}$$

This system is singular which means that the system determinant vanishes:

$$\begin{bmatrix}
< a_2a_2 > - < b_2b_2 > & < a_3a_2 > - < b_3b_2 > & < a_4a_2 > - < b_4b_2 > \\
< a_2a_3 > - < b_2b_3 > & < a_3a_3 > - < b_3b_3 > & < a_4a_3 > - < b_4b_3 > \\
< a_2a_4 > - < b_2b_4 > & < a_3a_4 > - < b_3b_4 > & < a_4a_4 > - < b_4b_4 >
\end{bmatrix} = 0$$

which is the view consistency constraint on four corresponding points in two views for calibrated orthographic projection cameras.

For the case of arbitrary unknown scale factors in the two views we have for four points:

$$\begin{bmatrix}
< a_2a_2 > -\rho < b_2b_2 > & < a_3a_2 > -\rho < b_3b_2 > & < a_4a_2 > -\rho < b_4b_2 > \\
< a_2a_3 > -\rho < b_2b_3 > & < a_3a_3 > -\rho < b_3b_3 > & < a_4a_3 > -\rho < b_4b_3 > \\
< a_2a_4 > -\rho < b_2b_4 > & < a_3a_4 > -\rho < b_3b_4 > & < a_4a_4 > -\rho < b_4b_4 >
\end{bmatrix} = 0$$

If this determinant is developed we get a polynomial in ρ. Denoting the determinants:

$$\begin{bmatrix}
< a_i\ a_2 > & < a_j\ a_2 > & < b_k\ b_2 > \\
< a_i\ a_3 > & < a_j\ a_3 > & < b_k\ b_3 > \\
< a_i\ a_4 > & < a_j\ a_4 > & < b_k\ b_4 >
\end{bmatrix} = [\ A_i\ A_j\ B_k\ ] \tag{17}$$

we get:

$$\begin{aligned}
&[\ A_2\ A_3\ A_4\ ] - \rho\ ([\ B_2\ A_3\ A_4\ ] + [\ A_2\ B_3\ A_4\ ] + [\ A_2\ A_3\ B_4\ ]) + \\
&+ \rho^2\ ([\ B_2\ B_3\ A_4\ ] + [\ B_2\ A_3\ B_4\ ] + [\ A_2\ B_3\ B_4\ ]) - \rho^3\ [\ B_2\ B_3\ B_4\ ] = 0
\end{aligned} \tag{18}$$

However, it is simple to show that:

$$[\ A_2\ A_3\ A_4\ ] = [\ B_2\ B_3\ B_4\ ] = 0 \tag{19}$$

The polynomial in ρ therefore reduces to:

$$\begin{aligned}
&[\ B_2\ A_3\ A_4\ ] + [\ A_2\ B_3\ A_4\ ] + [\ A_2\ A_3\ B_4\ ] - \\
&- \rho\ ([\ B_2\ B_3\ A_4\ ] + [\ B_2\ A_3\ B_4\ ] + [\ A_2\ B_3\ B_4\ ]) = 0
\end{aligned} \tag{20}$$

In order to eliminate the unknown scale factor ratio ρ we need a fifth point. We then get the system constraint:

$$\begin{bmatrix}
[B_2A_3A_4] + [A_2B_3A_4] + [A_2A_3B_4] & [B_2B_3A_4] + [B_2A_3B_4] + [A_2B_3B_4] \\
[B_2A_3A_5] + [A_2B_3A_5] + [A_2A_3B_5] & [B_2B_3A_5] + [B_2A_3B_5] + [A_2B_3B_5]
\end{bmatrix} = 0$$

which is a view consistency constraint for five points in two views of unknown scaled orthographic projection cameras

References

1. Aggarwal, J.K., Cai, Q., Human Motion Analysis: A Review, CVIU(73), No. 3, March 1999, pp. 428-440.
2. Baumberg A. and Hogg D. , Learning flexible models from image sequences, in: J.O. Eklundh, ed., Computer Vision-ECCV '94 (Third European Conference on Computer Vision, Stockholm, Sweden, May 2-6, 1994), Volume A, Springer, Berlin, 1994 299-308.
3. Bennett, B.M., Hoffman, D.D., Nicola, J.E., and Prakash, C., Structure from Two Orthographic Views of Rigid Motion, JOSA-A(6), No. 7, July 1989, pp. 1052-1069.
4. Bennett, B.M., Hoffman, D.D., and Prakash, C., Recognition Polynomials, JOSA-A(10), No. 4, April 1993, pp. 759-764.
5. Bregler C. Learning and Recognizing Human Dynamics in Video Sequences IEEE Conf. Computer Vision and Pattern Recognition, June 1997, Puerto Rico
6. Carlsson S. and Weinshall D., Dual computation of projective shape and camera positions from multiple images International Journal of Computer Vision Vol. 27 No 3, 1998
7. Cedras C. and Shah M. A survey of motion analysis from moving light displays, Proceedings, CVPR '94, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Seattle, WA, June 21-23, 1994), IEEE Computer Society Press, Los Alamitos, CA, 1994, 214-221.
8. Clemens D. and Jacobs D. Space and time bounds on model indexing *PAMI*, (13), 1007 - 1018 1991
9. Cutting J. and Kozlowski L. " Recognizing friends by their walk: Gait perception without familiarity cues," Bulletin of the Psychonomic Society, 9:253-356, 1977
10. Gavrila, D.M.], The Visual Analysis of Human Movement: A Survey, CVIU(73), No. 1, January 1999, pp. 82-98.
11. Hogg D. , A program to see a walking person, Image and Vision Computing, 1, (1):5-20, 1993
12. Johansson, G., Visual Motion Perception, SciAmer(232), June 1976, pp. 75-88.
13. Little J. and Boyd J.E. Recognizing People by Their Gait: The Shape of Motion Videre: Volume 1 • Number 2 Winter 1998
14. Niyogi S. A., and Adelson E. H. Analyzing and Recognizing Walking Figures in XYT Proceedings of Computer Vision and Pattern Recognition Seattle, WA; June (1994).
15. Polana R.and . Nelson R. , Recognition of nonrigid motion, Proceedings, ARPA Image Understanding Workshop (Monterey, CA, November 13-16, 1994), Morgan Kaufmann, San Francisco, CA, 1994, 1219-1224.
16. Quan L. Invariants of 6 points from 3 uncalibrated images, *Proc. 3:rd ECCV, pp. Vol. II* 459 - 470 1994
17. Rohr K. Towards model-based recognition of human movements in image sequences, CVGIP - Image Understanding, vol. 59, no. 1, 1994, 94-115.
18. Weinshall, D. Model-based invariants for 3D vision. *Int. J. Comp. Vision*, 10(1):27–42, 1993

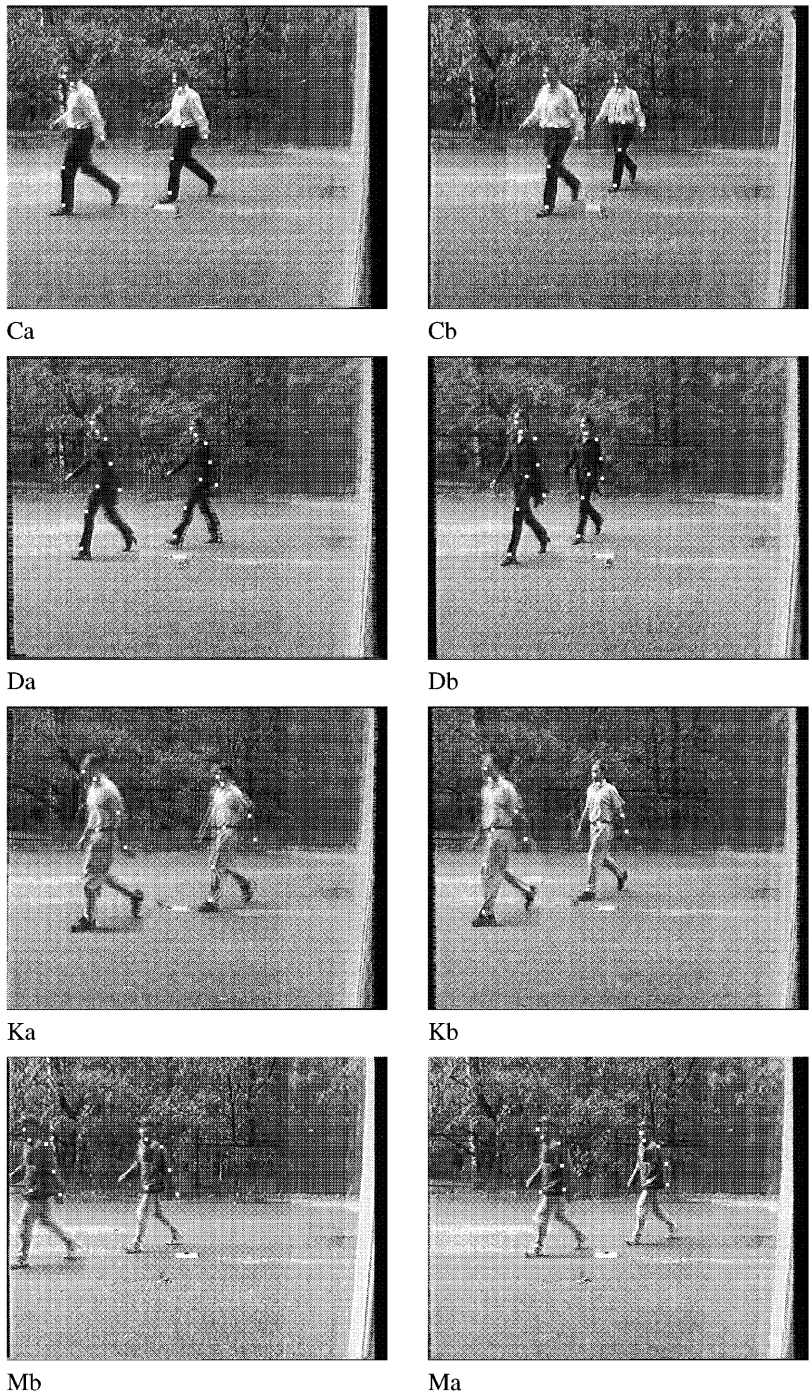

Fig. 10. First and last frames of 8 walking sequences with selected feature points