Available online at www.sciencedirect.com
SCIENCE DIRECT®
Clinical Biomechanics 20 (2005) 1–8
CLINICAL BIOMECHANICS
www.elsevier.com/locate/clinbiomech

Review

Autolabeling 3D tracks using neural networks

Stefan Holzreiter [a,b,*]

[a] Clinic of Orthopaedics in Ungulate, University of Veterinary Medicine, Vienna, Veterinärplatz 1, 1210 Wien, Austria
[b] Neurological Rehabilitation Centre NRZ-Rosenhügel, Rosenhügelstraße 192a, 1230 Wien, Austria

Received 16 April 2004; accepted 20 April 2004

Abstract

Motion capturing systems based on monochrome video have problems assigning measured 3D marker positions to the anatomically defined positions or labels of the markers applied to the test subject. This task is usually called "labelling" and is paramount to the reconstruction of 3D trajectories from a set of video frames from multiple cameras—the tracking procedure. Labelling means sorting a set of 3D vectors by their spatial positions. Neural networks can be made to "learn" from examples of marker positions in a given marker set, i.e. previously manually tracked video sequences. Trained neural networks are able to calculate a set of sorted approximate marker positions from an unsorted set of exact marker positions. The set of sorted exact positions can be found by pairing up both sets of marker positions via a minimum distance function. The neural network is trained only once and can then be applied to any number of individuals. The algorithm is designed for cyclic motions like for locomotion analysis.

Keywords: Autotracking; Autolabelling; Motion capturing

1. Introduction

Motion capturing of 3D movements is an important task in any laboratory for clinical gait or motion analysis. There exists a large variety of systems and different technical solutions for measuring the spatial movement of markers (Bhatnagar, 1993). Video based systems often preferred, because they are accurate and have robust and easily applicable wireless markers. A standard technique are sphere-shaped markers with a retroreflective surface and infrared spotlights coaxial to the camera lens. The retroreflective surface (usually made of prism reflex foil) reflects incoming light exactly back to the source. This property makes the markers appear as very bright spots on a black background in the video image, provided the exposure time is short enough. Other objects do not have retroreflective surfaces and will therefore appear much darker. Simple image processing software or even hardware may calculate a set of 2D co-ordinates of the marker locations within an image. The 3D marker-positions can be found by calculating the 3D intersection of the lines of sights of two or more cameras. For this calculation, the software must know the exact position of the cameras in its 6 degrees of freedom. These data are established in a calibration process prior to the measurement session (Abdel-Aziz and Karara, 1971; Kraus, 1996; Luhmann, 2000; Mikhail et al., 2001).

However, knowing the 3D marker-positions is only half the task. There will be several spots on the video and they have to be identified, i.e. assigned to their corresponding markers. The size of the spot cannot be used for identification since it depends on the distance of the marker from the camera lens. Markers with different shapes have other disadvantages. They must be large enough for the shape to be recognizable and the image processing routine is very complex, as the markers move and rotate and their 2D images change drastically depending on their 6 parameters (6 degrees of freedom) and illumination.

It should be easier and more reliable to identify a spot in 3D space (as belonging to a certain marker) from the spatial constellation of all spots in 3D space. It is an easy task for a human observer to recognize the individual

* Address: Clinic of Orthopaedics in Ungulate, University of Veterinary Medicine, Vienna, Veterinärplatz 1, 1210 Wien, Austria.

0268-0033/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.clinbiomech.2004.04.006

markers from a view of the complete constellation of all 3D positions if he knows where the markers are fixed on the subject and therefore may estimate all possible constellations during the measurement. Most video based motion capturing systems therefore include software to interactively identify 3D positions (Scheirman, 2003; Seeholzer, 2003; Brammall, 2003; Woolard, 1999). The procedure (usually called "tracking" or "labelling") is, however, time consuming and tedious work (Herda et al., 2000; Lopatenok and Kudrajashov, 2002).

Automation on the other hand, is not as simple as it may seem. An automatic tracking algorithm must deal with a number of exceptions. First, there is the case that a marker may temporarily be hidden. Second, there is the case that additional markers ("ghost markers") appear, if there are reflecting surfaces other than the markers in the measurement space. Ghost markers also appear if intersection of lines of sight are calculated by accident. Ghost markers may be reduced (but not entirely avoided) by setting a maximum distance between different lines of sight of a given marker. There is no universal algorithm for automatic labelling. Any practical algorithm will be restricted to a certain measurement configuration. Some parameters will always be needed to adjust the algorithm to such a configuration.

Commercial systems must consider different fields of application like motion capturing for the animation of movie characters, car crash tests, observations of mechanical parts or machines, biomechanical studies in sports, motion analysis of artists or motion analysis in a medical context (Aggarwal and Cai, 1999; Lopatenok and Kudrajashov, 2002; Moeslund and Granum, 2001). The usual autotracking programs therefore focus on applications with high commercial impact such as the movie business (Trager, 1999). However, the demands on the autotracking algorithms differ between applications in movie animations and clinical motion analysis. Movie animation needs to capture several individuals simultaneously. Medical motion-analysis, both for clinical and scientific purposes, always records one subject at a time, but requires many trials with different subjects under the same measurement setup. On the other hand, in most cases of medical applications some restrictions in the observed motions may be assumed. In locomotion analysis for example, motions are near cyclic and the subject moves in a straight line or on a treadmill.

Three of the world's leading manufacturers of video based motion-capturing systems, Motion-Analysis, Vicon and Peak-Motus (Motion-Analysis, Vicon and Peak-Moms are registered trade marks), use algorithms based on the distances between markers (Brammall, 2003; Woolard, 1999). Distances between markers on a single body segment (a segment can either be a rigid body or a chain of rigid bodies) can be assumed as approximately constant. The setup therefore needs a definition of the body segments and manually labelled example measurement data to gather the distances between the markers. To distinguish between segments with equal marker distances (for example: left and right leg, left and right arm) additional asymmetric markers are mounted on the subject to be measured. Anyway, when measuring a new subject, a new trial for "autolabel calibration" must be performed and manual labelling is necessary. Only the Peak-Motus system does not need this because it examines the proportional spacing among the markers and matches them to a reference template (Scheirman, 2003).

The algorithm presented here is not based on the distances between markers but the complete constellation of all measured 3D positions. Additional markers to distinguish between left and right limbs are not required. Between three and five manually tracked trials of different individuals are sufficient to teach the system. Subsequently, any number of different subjects can be tracked automatically as long as the same marker set is used and the motion pattern does not change drastically.

2. Terminology

The terms used to describe the reconstruction process of videometric data differ widely. What follows are brief definitions of terms as they are used in this paper:

- The term *spots* is used for the appearance of markers within the video images.
- *2D position* means the co-ordinates relative to the image.
- *3D positions* are spatial co-ordinates of an anonymous, unlabeled marker.
- The term *marker positions* is used for a 3D position assigned to a physical marker, i.e. a labelled 3D position.
- *Reconstruction*: Calculation of 3D positions from 2D image co-ordinates. The 3D position can be reconstructed by calculating the point of intersection of two or more lines of sight. The lines of sight can be derived from the previously measured spatial orientation of the camera ("calibration") and the 2D positions of the marker images gathered from the image analysis procedure. Errors may occur in the assignment of the rays to a specific markers, because the distance between the rays is the only available parameter for the assignment and it is never zero in practice.
- *Labelling*: Assignment of the anonymous 3D positions to the labels of the anatomically defined markers. Manual labelling is usual done for an arbitrary instant of time (a certain video frame). The extrapolation for the other frames is done per rectification.
- *Rectification*: Assuming that a marker cannot immediately jump from one position to another because

of the physically limited acceleration, the whole 3D trace of the marker can be reconstructed from one known initial position as long as the marker is not temporarily hidden. Rectification can be done by searching for the closest marker in the subsequent or previous frame or by extrapolation of the actual marker speed. Difficulties increase, as more markers disappear simultaneously. Normally, rectification needs a new initialisation (and therefore labelling) if the marker reappears after a certain hidden phase.

- *Interpolation*: Gaps in 3D tracks occurring if a marker was temporarily hidden may be reconstructed (or rather estimated) by linear interpolation or (more often) by spline interpolation.

3. The neural network

There are many different types of artificial neural networks (Köhle, 1990). The type used here can be characterized as an analog net with sigmoidal units using the backpropagation algorithm for training. Artificial neural networks are strongly simplified mathematical models of natural systems of neurons (Bishop, 1995; Duda et al., 2001; Rojas, 1993). Each neuron (also called "unit") has a number of inputs, a set of parameters called synaptic weights, and one output. A neural network consists of several neurons arranged in a structure of layers (one input layer, one or more hidden layers, one output layer). In case of a fully connected network as used in this application, the output of each unit is connected to each input of the subsequent layer. The way units are arranged and interconnected is called topology. In summary, the neural network can be considered to be a black box with n analog inputs and m analog outputs (Fig. 1).

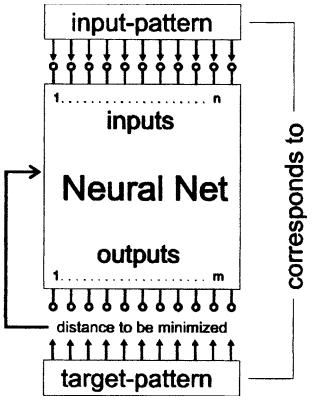

Fig. 1. The neural network in the training phase. Several input-patterns with corresponding target-patterns are presented to the network. The backpropagation algorithm tries to minimize the distance between the output and the target pattern by adjusting internal parameters (the synaptic weights).

Neural networks can (and must) be trained. During a training, a set of pattern data consisting of n-dimensional input patterns and corresponding m-dimensional target patterns are used to adjust the randomly initialised synaptic weights in an iterative process. This process uses the previously mentioned backpropagation algorithm (Rumelhart and McClelland, 1986) to minimize the differences between the output of the net and the target patterns. For detailed information about the structure and math of the neural network used here refer to (Holzreiter and Köhle, 1993).

After a successful training the net is able to perform an estimate of the target patterns by given input-patterns. Furthermore, if there exists a rule of how the target-patterns depend on the input-patterns, in most cases the network is able to derive this rule from the examples without an explicit knowledge of that rule and therefore make good output estimates from new input patterns as well. This feature is called prediction- or generalization-ability (Carney and Cunningham, 1999).

4. Preparation of data

Generally, practical applications of neural networks require some procedures for data preparation or feature preprocessing to reduce the amount of input data and to avoid disturbance by unimportant features (Jackson, 1997). For autolabelling, the most difficult task is to make the input data independent of the random permutation of the 3D position data. A special kind of data transformation based on so-called "virtual distance sensors" was therefore developed.

Nevertheless, some brief normalizations must be performed on the raw data first, in order to improve the learning success of the network. Fig. 2 gives an overview of the combination of the different normalizations and transformations applied during training and acquisition of the neural net.

First, the basic locomotion of the whole subject must be eliminated. This can be achieved by subtraction of the centre of gravity of all 3D positions. After this, the data should look as if the subject was walking on a treadmill.

Secondly, minimum and maximum co-ordinates values in each dimension must be calculated over the sequence and the tracks must be zoomed to fit a limited normalized space. This reduces the influence of subject size and simplifies subsequent calculations.

Finally (after the distance sensor transformation), the input and output signals of the neural network must be normalized, because their signal level is limited to range 0–1. Minimum and maximum values for each signal of the training pattern must therefore be retrieved and the signals must be transformed by a linear adaptation to a level of $\sim$0.25 for minimum and $\sim$0.75 for maximum. It

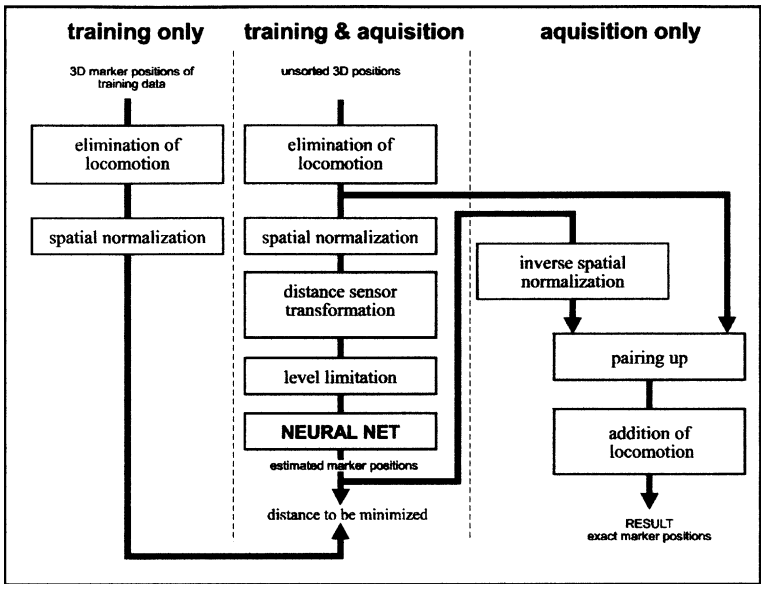

Fig. 2. Data flow of the autolabel algorithm. Left and middle columns show the procedures used during training of the neural network, middle and right columns show the parts used during data acquisition.

is important to have some reserve, because extreme values higher than the ones observed in the training data may occur.

5. The distance sensor transformation

The central feature of the autolabelling procedure is an algorithm for sorting a set of 3D positions and an inherent problem of any sorting algorithm is that input data appear in random order which overtaxes the learning capability of the neural net. A transformation of the unsorted 3D co-ordinates to the ordered input signals of the network is therefore required. The output of this transformation must depend on the constellation of the positions without the assumption of any sequence.

The initial idea for the distance sensor transformation was to divide the space of observation into large scale voxels in order to obtain a kind of marker-density distribution rather than spatial co-ordinate data. Subsequently, the straight borders of the voxels where made more fuzzy by calculating the sum of distances from the voxel centres weighted by a smooth distance function (closer markers have more weight). In detail, the space of observation is split into intervals of size 1 in each dimension resulting in a spatial arrangement of cubes of the size $1\times1\times1$. The number of intervals in each dimension can be set by parameter and depends on the proportion of the subject and the distribution of the markers within the space of observation. At the centre of each cube there is a virtual distance sensor (Fig. 3). The output signal of each distance sensor is the sum of the distances of all recorded 3D positions weighted by the Gaussian bell-shape function:

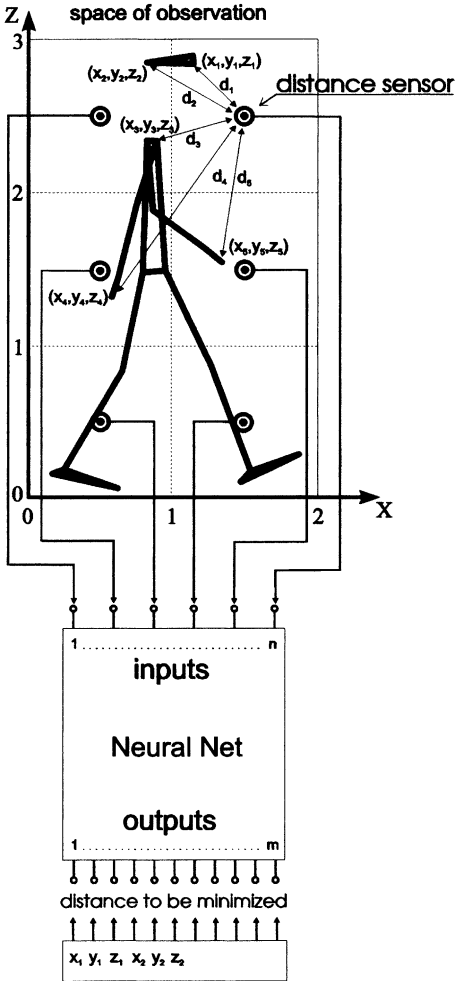

Fig. 3. The transformation from the set of unsorted 3D position into the input signals of the neural network is done by virtual distance sensors arranged in a spatial grid. As target pattern the true co-ordinates of the manually tracked marker-positions are used.

$$s = \sum \exp(-d_i^2)$$

s, output of the virtual distance sensor, d_i distance of the ith 3D position from the sensor.

6. Pairing up

The neural network can only estimate the positions of the sorted marker set. To achieve exact measurement results for each estimated position the corresponding actually measured position must be found. In cases of good approximation this can be done by searching the 3D position closest to the estimated marker-position. Unfortunately, this simple procedure may lead to double assignments. The perfect algorithm would have to try all permutations of the set of positions to find the one with the minimum sum of distances. As the number of permutations increases with $n!$, of the number of markers, computation time would be beyond limits of practicality.

A practical compromise is to avoid double assignment entirely by only taking vacant positions into consideration. Additional iterations trying to exchange the two worst assignments can be performed to reduce the sum of distances. Anyway, the assignment in pairs is a non-trivial optimisation problem.

The same algorithm can be used for rectification. Rectification means pairing up markers of subsequent frames either in crescent or descending order. The only additional feature required is the ability to handle disappearing and reappearing markers.

7. Results

A test-implementation (Holzreiter, 2003) of the autolabelling procedure was built based on a language system for motion analysis called MAL (Holzreiter and Jennings, 1996) and example tests have been performed with motion captures of walking or trotting horses plus rider on a treadmill (31 markers) and with human subjects in straight walk over a gaitway (21 markers).

The first test was performed at the Clinic for Orthopaedics in Ungulates of the University of Veterinary Medicine, Vienna, using a Motion-Analysis system for capturing and reconstruction.

The space of observation was split into three segments in x-direction (direction of movement) and two segments in y- and z-direction. The neural network consisted of 12 input synapsese ($=3\times2\times2$ distance sensors), 31 neurons in the hidden layer and 93 neurons ($=3$ dimensions$\times31$ markers) in the output layer. Five trials of 3 horses, each of them of about 10 s duration where used as training patterns and 500 000 iterations of back propagation where performed.

Fig. 4 shows the response of the network to input-data from a different horse (not part of the training-set) in 6 different instances of time. The highest deviation can be observed for the two markers on the head of the horse, because the motion of the head is often spontaneous and therefore difficult to learn for the net. Nevertheless, the pairing up algorithm finds the correct markers.

The second test (test with human subjects) was performed at the Neurological Rehabilitation Centre

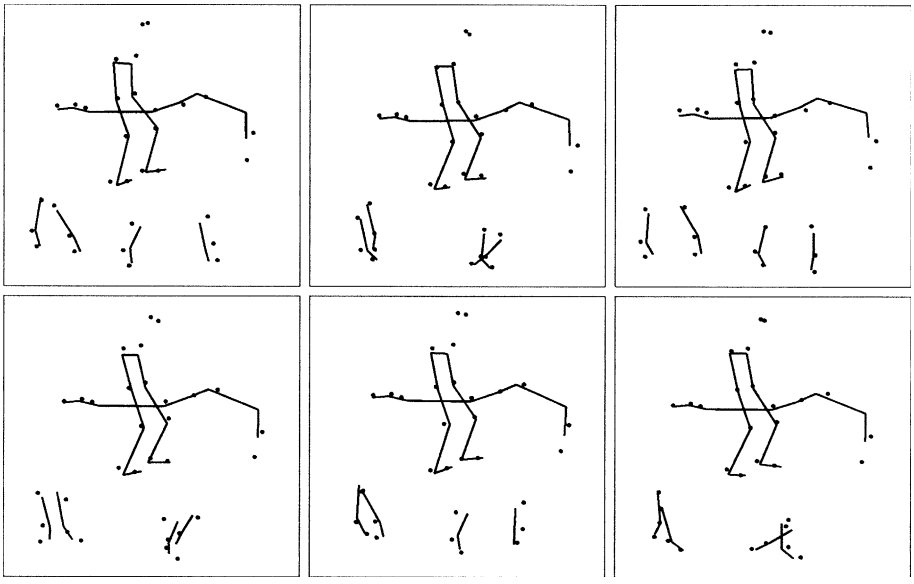

Fig. 4. Comparison of the (inserted, measured 3D positions of markers (drawn as circles) and the marker positions estimated by the neural network (stick-figure). The illustration shows six different instances in time of a horse with rider with a total of 31. The network was trained with patterns of three different horses but did not know the one shown in this figure. The data was acquired with a Motion-Analysis system at the Clinic for Ungulates of the University for Veterinary Medicine, Vienna.

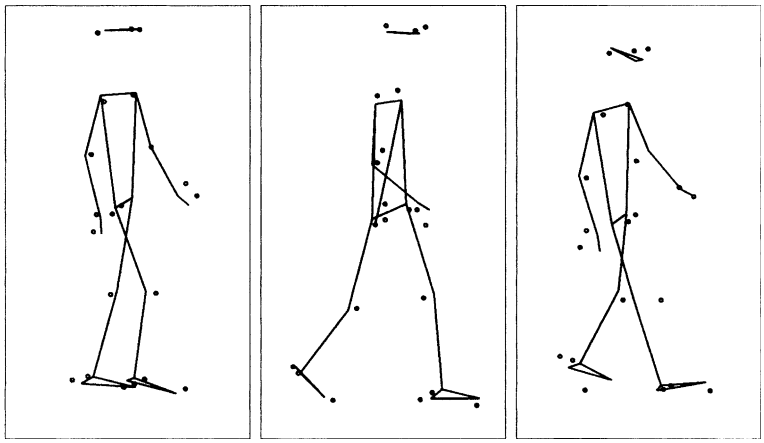

Fig. 5. Input and output values of the neural network of three different individuals, none of them part of the training set. The unsorted 3D positions (input data) are drawn as circles and the estimated positions of the sorted marker set (output data) are drawn as a stick-figure. The data was acquired with a Vicon 460 system at the NRZ-Rosenhugel, Vienna.

NRZ-Rosenhügel using a Vicon-460 system. Seven trials of three individuals walking on a straight walkway of 6 meters where used as the training set. The space of observation (after subtraction of the locomotion) was split into three segments vertically and two segments in each horizontal direction, resulting in 12 input synapsese for the net. Twenty-one units were used in *the* hidden layer and 63 units ($=3$ dimensions $\times 21$ markers) in the output layer. After a training of 1 500 000 iterations, the autotracking algorithm was tested with three further individuals. Fig. 5 illustrates the response of the neural network to three different subjects, none of them part of the training-set.

Some preconditions must be met when using the labelling algorithm:

- It works only if all markers (and no additional ghost markers) are visible within the frame.
- Motion must be comparable to the one of the training patterns.
- Markers must not be mounted very close to each other.

Anyway, the autotracking-procedure worked perfectly for several trials both of the first and the second series of tests as long as there were no dramatic disturbances in the input data like complete loss of markers (during the whole sequence) or marker positions drifting slowly away (this can not be recognized by the rectification algorithm).

8. The complete autotracking-procedure

As already mentioned, the autolabelling-algorithm only works for frames where all markers and no additional ghost-markers are visible. The autotracking-procedure therefore starts by counting the number of markers visible in each frame of the trial. It assumes that proper frames are found if the number of visible markers matches the actual number of markers. The complete sequence is then split into sections, each of them having a period of proper frames and they are limited by a local minimum of the marker-count function. Fig. 6 shows an example of a trial of a person with 21 markers walking on a gaitway. The first marker becomes visible approximately at frame 150 when the subject enters the measurement space. At the bottom of the figure the division into three sections can be seen.

The middle of the phase of suitable frames (marked with the characters "X") is taken as initial frame for applying the autolabel algorithm. The labelling of the previous and subsequent frames is done by rectification. Fig. 7 shows a complete sequence of tracked and auto-

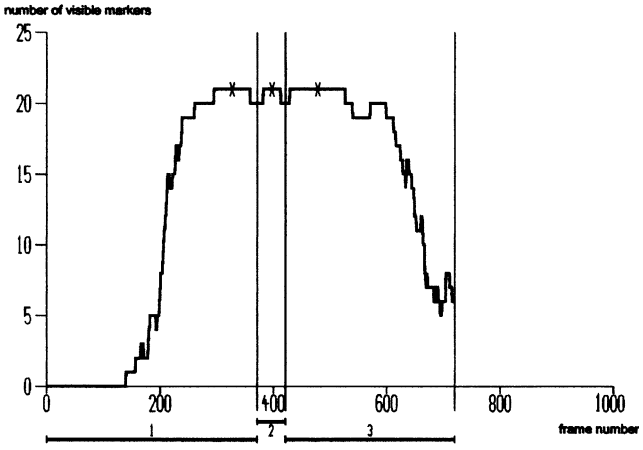

Fig. 6. Number of visible markers drawn against the frame number from a capture of a person with 21 markers walking on a gaitway. The first marker appears about in frame number 150 and the sequence stops about at frame 720 before the last marker leaves the measurement space. The sequence is split into 3 sections at local minima between phases of correct frames (with 21 markers). The characters "X" marks the frames taken for autolabelling.

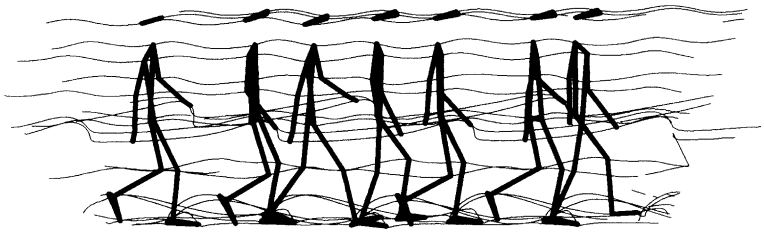

Fig. 7. The tracks of an automatically tracked sequence with inserted stick-figures at different instances of time. Rectification often fails at the very beginning and the end of a sequence when too many markers are hidden. However, in most cases this part of a sequence is of no interest anyway.

label led frames with stick figures inserted at certain instances of time to show the correct assignment of markers. Most difficulties occur at the very begin or end of the sequence where many markers are hidden because the subsect enters or leaves the space of observation.

9. Discussion

Some opportunities for farther improvements of the presented algorithm could be considered:

The neural network used here was a very simple one and the feature preprocessing and selection of the topology was done manually by trial and error. Modern techniques for optimising the performance of neural networks could be tested (Carney and Cunningham, 1999). Cross-validation e.g. is a technique to select the best learning success from various trained networks and can be used to optimise the network topology, too (Fiesler, 1997). Backpropagation is a simple but slow method for training. Faster and more effective methods could be suggested. Furthermore, there are methods to assess the quality of the feature preprocessing (Egmont-Petersen et al., 1998).

An important improvement would be to enable the algorithm to work under the presence of hidden markers or ghost markers. Rectification would become superfluous and more reliability specifically at the beginning or end of a sequence (when the subject enters or leaves the space of observation) could be expected. Unfortunately, the present algorithm is very sensitive to superfluous or missing markers.

10. Conclusion

Artificial neural networks are a new way to assign 3D tracks to marker labels. The most important advantage compared to algorithms based on distances between markers is the ability to track different subjects with a single setup. Furthermore, the setup only needs some sample data of manually tracked trials and a few parameters (the number of distance-sensors and the topology of the neural net). It does not need a detailed list of body segments and no additional markers to distinguish among symmetric segments are required.

Disadvantages are the restriction to cyclic motion patterns and the inability to track multiple subjects captured simultaneously. The latter one might be of minor importance for most clinical and scientific applications. It is expected that there exists a limitation for highly pathologic motion patterns, but they have not been tested yet. A well prepared measurement environment and high quality equipment is required to obtain clean video data and a minimum of ghost markers.

Particularly, applications where a large number of subjects must be tracked, could benefit from this method.

References

Abdel-Aziz, Y.I., Karara, H.M., 1971. Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. In: Proceedings of the Symposium on Close-Range Photogrammetry. American Society of Photogrammetry, pp. 1–18.

Aggarwal, J., Cai, Q., 1999. Human motion analysis: a review. Computer Vision and Image Understanding 73 (3).

Bhatnagar, D.K., 1993. Position trackers for Head Mounted Display systems: A Survey. Technical Report TR93-010, University of North Carolina at Chapel Hill.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

Brammall, A., 2003. Real Time Optical Motion Capture Systems from Motion Analysis Corporation [Internet], Motion Analysis Corporation. Santa Rosa, California. Available from <http://www.motionanalysis.com/index.html> [Accessed 28 May, 2003].

Carney, J., Cunningham, P., 1999. The NeuralBAG algorithm: optimizing generalization performance in bagged neural networks. In: ESANN'1999 Proceedings, European Symposium on Artificial Neural Networks Bruges (Belgium). D-Facto Public., pp. 135–140. ISBN 2-600049-9-X.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. John Wiley & Sons.

Egmont-Petersen, M. et al., 1998. Assesing the importance of features for multi-layer perceptrons, neural networks 11. Pergamon. pp. 623–635.

Fiesler, E., 1997. Neural Network Topologies [Internet] Handbook for Institute of Physics. IOP Publishing LTD. Available from <http://www.iop.org/Books/CIL/HNC/pdf/NCB2.PDF> (Accessed 3 August 2003).

Herda, L., et. al., 2000. skeleton-based motion capture for robust reconstruction of human motion. In: Computer Animation, Philadelphia, USA.

Holzreiter, S., 2003. MAL Dokumentation: Videometrie [Internet]. Available from <http://www.8ung.at/holzreiter_mal/videomet.htm> (Accessed 3 August 2003).

Holzreiter, S., Jennings, S., 1996. Programming language for motion analysis. Human Movement Science 15, 497–508.

Holzreiter, S., Köhle, M., 1993. Assessment of gait pattern using neural networks. Journal of Biomechanics 26 (6), 645–651.

Jackson, T.O., 1997. Data Input and Output Representations [Internet] Handbook for Institute of Physics. IOP Publishing LTD. Available from <http://www.iop.org/Books/CIL/HNC/pdf/NCB4.PDF> (Accessed 3 August 2003).

Köhle, M., 1990. Neurale Netze. Springer-Verlag, Wien.

Kraus, K., 1996. Photogrammetrie, Band 2.3. Auflage. Dümmler Verlag, Bonn. p. 105.

Lopatenok, T., Kudrajashov, O., 2002. The model-based approach of markers identification and visualisation in motion capturing systems. In: Simulation und Visualisierung 2002, Otto-von-Guericke-Universität Magdeburg.

Luhmann, T., 2000. Nahbereichsphotogrammetrie. Herbert Wichmann Verlag, Heidelberg. p. 235.

Mikhail, E. et al., 2001. Introduction to Modern Photogrammetry. John Wiley & Sons Inc., NY, USA. p. 251 (Chapter 9.3.2).

Moeslund, T., Granum, E., 2001. A survey of computer vision-based human motion capture. Computer Vision and Image Understanding 81, 231–268.

Rojas, R., 1993. Theorie der Neuronalen Netze. Springer, Berlin/Heidelberg/New York/Tokyo.

Rumelhart, D., McClelland, J., 1986. Parallel Distributed Processing. MIT Press, Cambridge, MA.

Scheirman, G., 2003. Peak Performance Technologies, Inc. Home [Internet], Peak Performance Technologies Inc. Available from <http://www.peakperform.com/> (Accessed 17 July, 2003).

Seeholzer, T., 2003. SIMI°Motion Gait 3D [Internet] Simi Company. Available from <http://www.simi.com/en/produkte/motion/gait.html> (Accessed 26 May 2003).

Trager, W., 1999. A Practical Approach to Motion Capture: Acclaim's Optical Motion Capture System [Internet] Advanced Technologies Group. Available from <http://www.siggraph.org/education/materids/HyperGraph/animation/character_animation/motion_capture/motion_optical.htm> (Accessed 4 August 2003).

Woolard, A., 1999. Vicon ST2 User Manual. Oxford Metrics Ltd. pp. 86–91 (Chapter 3.4).