Scale-Based Human Motion Representation for Action Recognition

Janez Perš[1], Goran Vučković[2], Branko Dežman[2] and Stanislav Kovačič[1]

[1]Faculty of Electrical Engineering
University of Ljubljana
Tržaska 25, SI-1000 Ljubljana, Slovenia
{janez.pers},{stanislav.kovacic}@fe.uni-lj.si

[2]Faculty of Sport, University of Ljubljana
Gortanova 22, SI-1000, Ljubljana, Slovenia

Abstract

The design of action recognition algorithms often relies on knowledge of the particular problem, which is not always available. Moreover, algorithms usually incorporate a number of parameters, which influence their performance. To solve these problems, we explore the possibility of developing more general action recognition algorithms by systematic reduction of complexity of human motion, instead of designing more and more complex algorithms. We see the key to reducing complexity in systematic decomposition of human motion to different scales, each representing different level of motion detail. We assume that human actions influence different scales of motion, and they should be observed that way. The features, obtained on different scales of motion can be joined together to represent the complex motion in uniform and manageable way. Our approach was tested in the sports domain, on a particular problem of detecting the action of athlete hitting the ball with a racquet in the game of squash. Video recordings of actual tournament match were used, and manual annotations were provided by squash expert as a ground truth.

1. Introduction

Analysis of human motion is a challenging problem mainly because of its complexity. Many researchers work in the field of computer vision based human motion analysis. This is reflected in several surveys on this topic [1, 6, 4], covering various areas of the field. Important area of human motion analysis is recognition and detection of human actions, with wide range of applications, from automatic annotation of sports video to the fully automatized security systems. However, design of action recognition algorithms often relies on detailed knowledge of the particular problem, which is not always available. Moreover, algorithms usually incorporate a number of parameters, which influence their performance. Many researchers [3, 2, 11, 12, 7] report high recognition rates for their particular action recognition problems. These solutions employ complex and sophisticated algorithms, which match the complexity of human motion to the degree needed for particular problem. The complexity of solutions however raises the question of generality - namely, how deeply are the actual problem and its specifics rooted in the proposed solutions? Number of parameters, hidden in these algorithms is usually large and the systematic analysis of their influence is nearly impossible.

In this paper, we explore the possibility of developing more general action recognition algorithms by systematic reduction in complexity of human motion. We see the key to reducing complexity in systematic decomposition of human motion to different scales, each representing different level of detail. In this way, complex human motion can be expressed as a combination of many simpler components, which are clearly visible only when looking at the right scale. The underlying assumption of presented work is that human actions influence different scales of motion, and should be observed that way. We argue that the proper decomposition of human motion results in a number of simpler problems that can be addressed with more general approaches. The features, obtained on different scales of motion can be joined together to represent the complex motion in uniform and manageable way.

Not *every* action recognition problem needs multiscale observations of human movement and we do not strive to design a *general* action recognition algorithm, which will be a difficult task for many years to come. However, we believe that our approach will lead to faster application of developed methods to real-world problems, which do not

require 100% recognition rates, such as sports video indexing, abstracting and highlighting. These areas would benefit tremendously from automated video annotation systems, however, they require simple and stable algorithms, given the diverse nature of video recordings.

This paper is structured as follows: first, we present the scale-based representation of human motion which allows systematic classification of various human motion analysis techniques. We present the reasons for choosing sport match as our testbed. The next section describes our implementation of scale-based decomposition of human motion: the simple tracking algorithm and the analysis of human motion using two complementary motion recognition algorithms, which operate on two different scales of motion. Before concluding, we show that the fusion of both sets of features improves the recognition results.

2. Scale-based human motion representation

Classification of video-based human motion analysis techniques is not uniform and largerly depends on interests of a particlular author [1, 6, 4]. There are some common points, for example the division to two large areas of motion analysis (analysis of whole body motion vs. analysis of motion of the body parts), since these two problems are seen as fundamentally different.

Such classification forms a basis for definition of human motion scale. Analysis of whole body motion looks at the human on the large (coarse) scale, essentially representing its position with a single point. Tracking of the body parts looks at a human at smaller (finer) scale, looking for details. Analogy with the classical scale-space [13] example of ”looking at the trees” vs. ”looking at the forest” is obvious.

Scale-space representation of the world asserts that some properties of the observed object appear only when observed at a proper scale. In scale-space, every observation has additional parameter - scale δ. However, human movement is a complex spatio-temporal phenomenon which has at least two spatial and one temporal dimension and the scale of observation is defined by a number of parameters - resolution, sampling rate, width of observation windows and similar. One possible definition of human motion scale is shown in Fig. 1. In the real world, the observer never sees the full scale of the motion. The visible interval of scale is determined by camera setup and geometry (zooming in reveals finer scales of motion) and the sensor resolution - observer cannot see the details that are below the resolution of the CCD chip or are faster than video acquisition frame rate.

Such representation of human motion offers possibility for classification of human motion analysis methods, based on the observed object - human body, regardless of their actual implementation. For example, whole-body tracking of people in a parking lot for security purposes operates mostly in the coarsest part of the scale, observing just statistical properties of human motion. Tactical sport analysis operates on a slightly finer scale, recording positions and velocities of players during the match. Gesture recognition observes even finer motion of arms and fingers. Video-based human analysis algorithms essentially try to focus on the desired interval of the scale, which provides the most usable information. The structure of algorithm and its parameters determine the interval of scale that is visible to them.

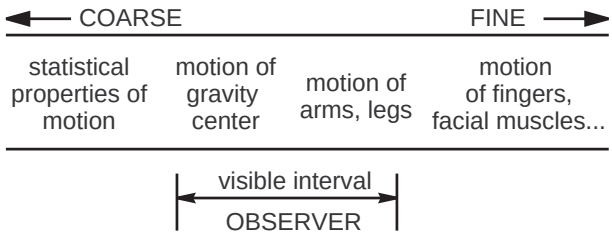

Figure 1. Human motion scale as seen from computer vision perspective.

2.1 What is the right scale?

In the case of tracking, the scale of interest should be the basic specification, determined even before the algorithm is developed. Different tracking methods are focused to different scales, as demonstrated in [10]. They may be very accurate within its scale of motion, however if the observation scale is wrong for the particular application, they will be labeled as inaccurate.

In action recognition, the issue of the right scale becomes more difficult. Most of the algorithms for action recognition focus on the single scale. For example, [2] operates on the trajectories of several larger body parts. Slightly finer scale (movement of arms) is observed in [11]. Sometimes, algorithms explicitly normalize video sequences around the perceived body center to suppress whole-body motion [7, 12] and discard the trajectories. Field of motion vectors across the whole human body is observed in [3], capturing wider range of scales, although their individual contributions to action detection are not clear. We may ask, do actions really influence only narrow parts of the human motion scale?

We built on the hypothesis, that *human actions and activities generally reflect on more intervals of the human motion scale.* For example, threat by street robber may be exhibited as a sudden move, threatening gesture and appropriate facial expression. We also think that *the motion on the different scales is uncorrelated, unless it is influenced by certain action or activity.* Walking person will exhibit motion of its

body center and its body parts. The presence of only body center motion would suggest some other activity (perhaps riding the moving walkway?)

The right scale for human action recognition should be defined as the scale or *the scales* that contain useful information for action recognition. For many action recognition problems we simply do not know which scales carry information. Therefore, for the recognition system to be general, it should learn these in some automatic way. Our concept is to build a system from several algorithms, focused on different scales, and then join the resulting information.

3. Input data

Our application domain is a squash match. Squash is an indoor racquet sport, played on a well illuminated 9.75×6.4 m court by two athletes. Player wins the match by winning three sets. As most of the sports, squash has well defined rules and long record of research focused on player movement. The most important action of a player is hitting the ball with a racquet before it hits the ground. Cooperation with the sports experts enabled us to obtain digital video of a tournament match, along with annotations describing the exact moment and type of a hit. One frame of video, recorded at 384×288 pixel resolution and 25 frames per second is shown in Fig. 2.

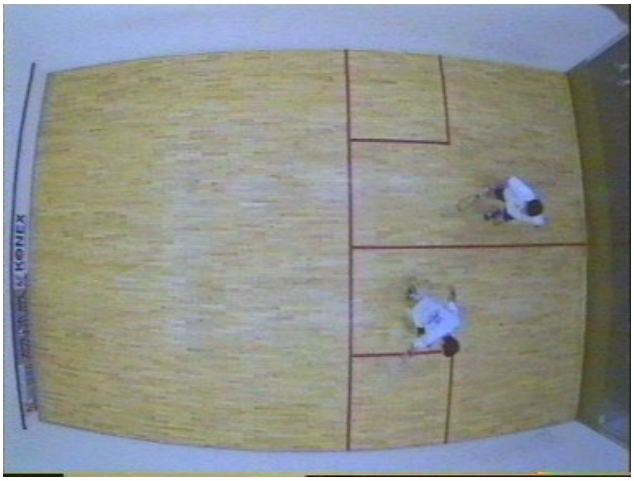

Figure 2. One frame of input video.

Detecting hits in squash match represents a well defined real-world problem, as such annotations are needed for match analysis. Fig. 3 shows a sequence of frames in region of interest of a single player. The problem is however difficult, since actions we are to detect are not dominant part of the video.

Camera was calibrated and simple background subtraction algorithm was used to obtain player trajectories. Tracking of players was done under supervision of the sports expert, who had the ability to stop the tracking and correct the obtained positions. The only parameter of the tracking was the desired number of pixels above the threshold in difference images, which depends only on number of players in the scene. Algorithm corrected the threshold by one after each frame was processed, based on the number of detected positive pixels. Median filtering of thresholded image was performed to obtain blobs. Their centers of gravity were used as player positions. Players positions on previous frame were used to separate merged blobs when players collided. Such system is used for trajectory analysis by the sports experts, and has been found to work well across several tournaments. Video data, trajectory data (in the court coordinate system) and expert annotations formed the testbed for our action recognition algorithm.

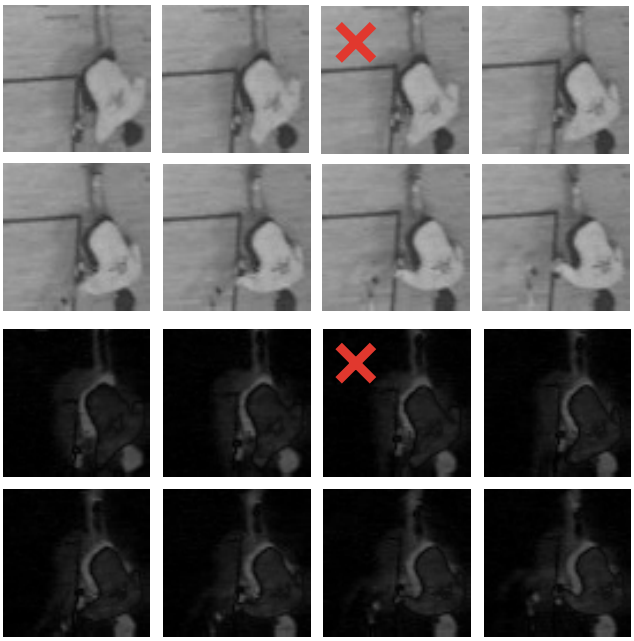

Figure 3. Image sequence of player, hitting a ball. Rows 1 and 2 show grayscale images, rows 3 and 4 show the images with background subtracted. Image, marked by X was annotated by expert as the exact moment of hit.

4 Action recognition

Our aim is to build action detection system, which would process the continous stream of video, and detect the moment when interesting action takes place. To develop and test the approach, the intervals of both trajectory and video were extracted and divided into the two classes: ω_1 - player is hitting the ball, and ω_2 - player is *not* hitting the ball. First

class is defined by the expert annotations (we used all types of hits except serves), and the second class was sampled at the middle of intervals between the hits plus one arbitrarily chosen sample, to obtain the classes with same number of samples. This way, the problem was transformed to the problem of classification between ω_1 and ω_2. It should be noted that the apriori probability of ω_1 is low (below 5%, if we observe hits through 5 frame window), which we did not take into the account. The training of the algorithm was done on one set (S) of the match (19,980 frames, 158 actions), and the testing was done on another (test set T - 18,344 frames, 148 actions). Only one player was observed.

4.1 Obtaining the parameters

Trajectories and image sequences contain information about coarse and fine scale of player motion, respectively. We expected that trajectories will exhibit distinctive shapes, and that image sequences would capture the movement of player body parts in process of hitting the ball. Different nature of these two types of input data requires different processing. The choice of algorithms places each of the methods at its place within the scale-space of human movement (Fig. 1), while the parameters of algorithms limit the chosen scale more accurately. The right values of parameters for certain task are difficult to guess, therefore we employed exhaustive search within the reasonable range and accepted the parameters which gave highest recognition rates. The whole procedure can be outlined as follows:

1. Split the annotations from training set S into the subsets S_1 (75%[1]) and S_2 (25%). Initialize parameters.
2. Based on annotations, extract the segments from input data according to current parameter set. This defines the scale!
3. Train the classifier on S_1.
4. Test the classifier on S_2, save the results, and return to 2, unless the search space is exhausted.
5. Accept the parameters with highest recognition rate for both classes.

4.2 Training and classification

To classify the samples, we used Linear Discriminant Analysis (LDA) [9]. Given the training set of m classes, LDA provides the transformation matrix $\mathbf{W}_{LDA}$, which transforms the input vectors into the $m-1$ dimensioned feature space, to ensure best classification. In our case ($m = 2$), we obtained one value for each sample. $\mathbf{W}_{LDA}$ was calculated on S_1. The decision threshold between ω_1 and ω_2 was found using the nonlinear maximization (Nelder-Mead simplex, Matlab FMINS) of total classification rate on S_1.

[1]Since we use S_1 to compute PCA or LDA representation of the class, the number of samples in S_1 has to be as large as possible.

4.3 Coarse scale - the trajectories

Three parameters control the extraction of the samples from the continuous trajectory. Trajectory smoothing reduces noise *and* details in the trajectory. The amount of smoothing is defined by width of the Gaussian smoothing kernel (extending from -3σ to $+3\sigma$), which is the same for x and y component. The width of observation window is the second parameter. The third parameter is the delay between the moment of annotation and the center of the observation window. It takes into the account the fact that action may resonate on different scales with some delay. The samples of trajectories were normalized after the extraction to suppress the coarser parts of the motion scale (absolute position and rotation). First, their mean value (for x and y component) was normalized to zero. Their mean rotation around the zero point was subsequently also normalized to zero. x and y parts of trajectory were concatenated to single vector before feeding them into the LDA.

The normalized trajectories for optimal set of parameters are shown in Fig. 4. To help the reader visualize the difference, manually sketched shapes are placed right to the trajectories. Trajectories in the ω_1 class (hits) exhibit more bending (U-shape) than those from ω_2. This is consistent with the squash game - player will approach the ball, hit it with the racquet and retreat to make space for the other player. The classes overlap for significant amount - the recognition rate for the test set T was 66% for ω_1 and 70% for ω_2. The impact of parameters on recognition rate for

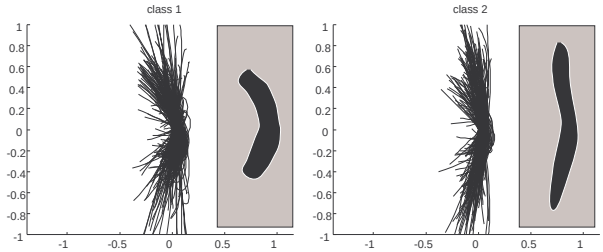

Figure 4. Normalized trajectories for the train set S.

S_2 is shown in Fig. 5. Graphs have been obtained by iterating one parameter through the search space and keeping the others at their optimum position. The search space and optimum set can be inferred from the presented graphs. The whole iterative process took about 20 minutes in Matlab on Intel Pentium II 400 MHz processor.

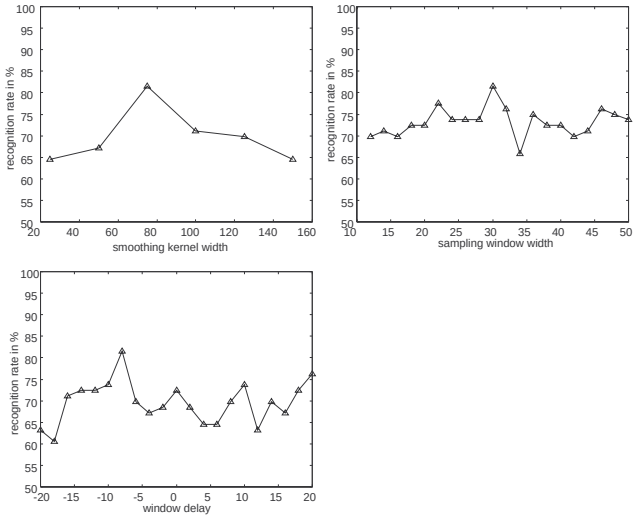

Figure 5. Impact of parameters (first scale) to the total recognition rate of S_2. All values are in frames (1/25 s).

4.4 Fine scale - the images

Four parameters control the extraction of the samples from the image sequence. First, the sequence of difference images was generated by subtracting each frame of video from the image of the empty court. This was necessary, as majority of hits take place near the wall, and algorithm tends to learn the walls instead of action. The difference images are shown in the last two rows of Fig. 2. The trajectories, smoothed with the kernel of fixed width were used to extract windows of size 80×80 pixels from the sequence. More general approach would require inclusion of smoothing kernel width into the parameter set, however, we relied on our previous experiments to select the width of 25 samples. The centers of image (pixel) gravity were calculated and images were aligned, such that center of image corresponded to the gravity center, to compensate possible inaccuracies in trajectories.

The images were cropped to the size, defined by the first extraction parameter and arranged into sequences, with sequence length being the second parameter. The delay between the annotation and the center of sequence is the third, the downscaled size of (already windowed) image is the fourth parameter. Downscaling has been performed by bilinear interpolation. It has the two benefits: it allows the algorithm to run faster and supresses the finer part of the motion scale. All pixels of each image sequence were then arranged into vectors of length n. In contrast to first scale, we cannot apply LDA on such vectors. By definition of LDA, we need at least n training samples, otherwise computational difficulties occur. The solution is to use PCA [9]. This is known concept of "eigensequences", as described by [8].

We calculated PCA transformation matrix $\mathbf{W}_{PCA}$ from the samples from class ω_1. The actions are not the dominant part of the sequences, and we cannot use the eigenvectors that correspond to the first few largest eigenvalues (PCA by itself provides optimal signal *representation*, not *classification*, [5]). Our tests have shown that these capture very little temporal action, which is visible only in lower-valued eigenvectors. The solution [9] is to apply LDA on top of the PCA-transformed features, to automatically extract those features that contribute most to classification. The dimension of the intermediate space was limited to 119 (number of train samples less one) vectors due to computational implementation of the PCA. A quick look at the $\mathbf{W}_{LDA}$ revealed that LDA indeed favoured lower-valued eigenvectors. The results of classification of the testing set T were 72% and 81% for ω_1 and ω_2, respectively. The impact of parameters on recognition rate for S_2 is shown in Fig. 5. Graphs have been obtained in the similar way than those in Fig. 4. It should be noted that the process of iteration takes considerably longer for this scale - 10 hours in Matlab on AMD Athlon 1 GHz processor. The main reason for this is inclusion of PCA in the iterative loop.

Both Fig. 4 and Fig. 5 show that some of the parameters introduce distinct peaks in recognition rate, and others do not. The procedure of searching through the parameter space is usable to obtain the *working subspace* of parameters when the details of motion and actions are not available.

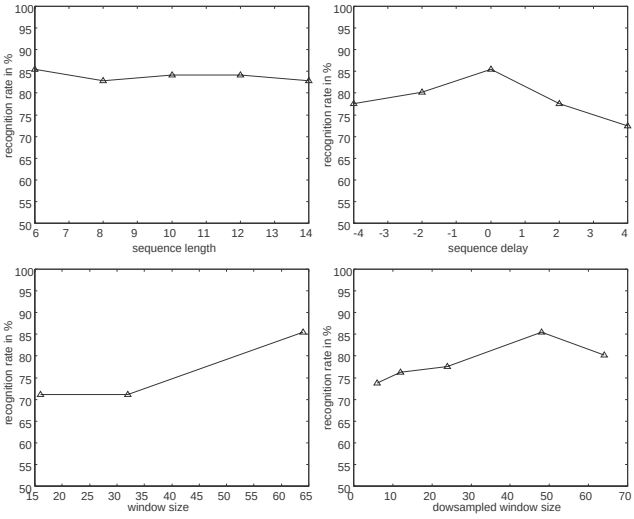

Figure 6. Impact of parameters (second scale) to the total recognition rate of S_2. Values are in frames (1/25 s) and pixels.

4.5 Recognition on both scales

Our main assumption was that actions reflect themselves on different scales of human motion. To obtain better results, we joined the information from both scales. The test set T was transformed to the feature space, using algorithms for both scales, trained on the train set S. Learned decision boundaries for both algorithms were discarded, and both results were joined to form two dimensional vectors, one dimension for each scale. Fig. 7 shows the samples in this 2D space.

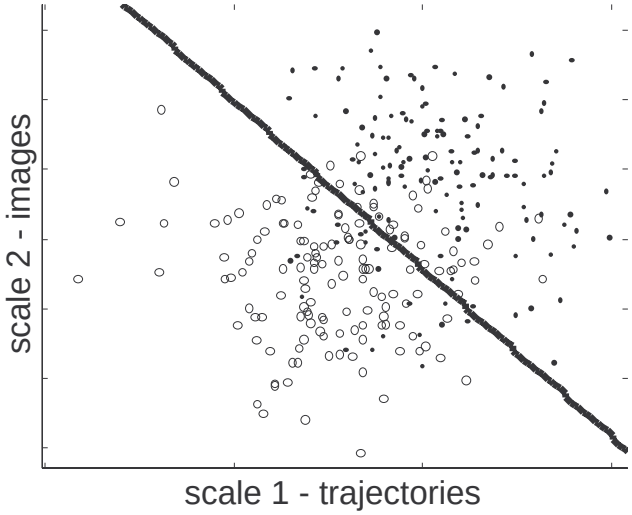

Figure 7. Clusters and decision boundary. ω_1 - circles, ω_2 - dots.

Clusters are still overlapping, however, it can be clearly seen that they are placed along the diagonal. This shows that the information from both scales works complementary. Test on this set of data has been performed by randomly selecting 20% of the samples for learning decision boundary by the means of Linear Least Squares method and testing the classification on remaining 80% (the holdout method). The test was repeated 100 times, and recognition rates were 82% for ω_1 and 81% for ω_2.

5 Conclusion

Although the resulting recognition rates are lower than in some of the reported work, they illustrate the importance of observing the human actions on the wider spectrum of scales, especially when facing such difficult problems. In our case, actions are not well recognizable when looking at one scale alone, since in some instances they simply may *not* result in particular trajectory or particular type of body motion. This problem has nothing to do with classifier design (since the information may simply not be there). We expect that many real world problems will exibit such problems once closely examined, and that the proper way of addressing them is by fusing together information from many motion scales.

Despite the low recognition ratios, such solution could still help in some limited tasks. Used in video annotation, it would for example reduce the amount of video data for nearly five fold, and those 20% of video would still contain 80% of the actions, given the apriori probabilities of both classes.

References

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. In *IEEE Nonrigid and Articulated Motion Workshop*, pages 90–102, Puerto Rico, June 17-19 1997.

[2] A. Ali and J. K. Aggarwal. Segmentation and recognition of continuous human activity. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 28–35, Vancouver, Canada, July, 8 2001.

[3] B. A. Boghossian and S. A. Velastin. Image processing system for pedestrian monitoring using neural classification of normal motion patterns. *Measurement and Control (Special Issue on Intelligent Vision Systems)*, 32(9):261–264, 1999.

[4] I. A. Essa. Computers seeing people. *AI Magazine*, 20(1):69–82, 1999.

[5] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, London, 2nd edition, 1990. Chapters 9 and 10.

[6] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, 1999.

[7] N. Krahnstover, M. Yeasin, and R. Sharma. Towards a unified framework for tracking and analysis of human motion. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 47–54, Vancouver, Canada, July, 8 2001.

[8] N. Li, S. Dettmer, and M. Shah. Visually recognizing speech using eigensequences. In M. Shah and R. Jain, editors, *Motion-Based Recognition*, pages 345–371, 1997.

[9] A. Martinez and A. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, February 2001.

[10] J. Perš, M. Bon, S. Kovačič, M. Šibila, and B. Dežman. Observation and analysis of large-scale human motion. *Human Movement Science*, 21:295–311, 2002.

[11] C. Rao and M. Shah. View-invariant representation and learning of human action. In *IEEE Workshop on Detection and Recognition of Events in Video*, pages 55–63, Vancouver, Canada, July, 8 2001.

[12] R. Rosales and S. Sclaroff. 3d trajectory recovery for tracking multiple objects and trajectory guided recognition of actions. In *CVPR 1999*, Fort Collins, Colorado, June 23-25 1999.

[13] J. Sporring, M. Nielsen, L. Florack, and P. Johansen, editors. *Gaussian Scale-Space Theory*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.