Computer Vision — CVWW'03, Ondřej Drbohlav (ed.)
Valtice, Czech Republic, February 3–6, 2003
Czech Pattern Recognition Society

Human Activities at Different Levels of Detail

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

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

[2]Faculty of Sport, University of Ljubljana
Gortanova 22, SI-1000, Ljubljana, Slovenia
goran.vuckovic@sp.uni-lj.si, brane.dezman@sp.uni-lj.si

Abstract *Classification and recognition of human activity is an important field od computer vision based human motion analysis. The main problem of human motion analysis is the great complexity of human motion, which is rooted in the complicated structure of human body itself. In this paper we explore an alternative view of human motion analysis, which is based on the concept of human motion scale. This concept allows analysis of important issues in computer vision based human motion analysis that have not been sufficiently explored in the related work. In this paper we focus on one particular issue: the appropriate level of detail for a particular human activity classification task. For our experiments we used a video database of recorded squash matches, which have been processed using simple blob tracker under human supervision to obtain positions of players during the whole match. Digitized recordings were annotated by sports experts to provide data about two types of player activity. Annotations are used as a ground truth in our experiments. Results confirm that the choice of the level of detail on which human motion is observed may play crucial role in the efficiency of the classification algorithm. Additionally, we show that our findings are consistent over the whole database of 18 sets of squash play, which were recorded at different tournaments with different players.*

1 Introduction

Analysis of human motion is a challenging problem mainly because of complexity of the motion. Many researchers work in the field of computer vision based human motion analysis. This is reflected in several surveys on this topic [1, 5, 4, 11], covering various areas of the field. The field of human motion analysis can be roughly divided into two areas: human motion acquisition (tracking) and classification, recognition and detection of human activity. Although most of the algorithms that deal with human activity classification work on top of the various tracking methods, this is not necessary the case, as in [13]. Possible applications of human motion analysis algorithms range from automatic annotation of sports video, human-machine interaction to the fully automated security systems. Many researchers [3, 2, 8, 9, 6] report high recognition rates for their particular activity recognition problems. These solutions employ complex and sophisticated algorithms, which match the complexity of human motion to the degree needed for particular problem.

However, there are many issues that have to be resolved before such systems are put into widespread use. Since most of the human tracking algorithms actually *measure* human motion, it is surprising that most of the researchers did not perform adequate quantitative analysis of measurement errors. When such analysis was performed, as for example in [7], researchers quickly discovered the problem of ground truth definition, which is closely related to the complex nature of human motion. When real-world applications are to be developed, the apparently simple questions, such as *What does it mean "running"? What exactly is "standing still"?* become much more difficult to answer.

These questions implicitly require the definition of *observation scale*, on which particular method observes human motion. In the next part of the paper, we will define the concept of *human motion scale* and show how it relates to the already established classification of human motion analysis algorithms. In the central section of our work, we will illustrate our hypothesis that the choice of the right scale may play a vital role in the performance of a simple activity classification algorithm, using the sports domain data. Before concluding the paper, we will present results, which show that our findings are consistent over the large database of human motion data.

2 Human motion scale

Classification of video-based human motion analysis techniques is not uniform and largerly depends on interests of a particlular author [1, 5, 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 [10] 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. It should be noted, however, that 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. The definition of human motion scale is shown in Figure 1.

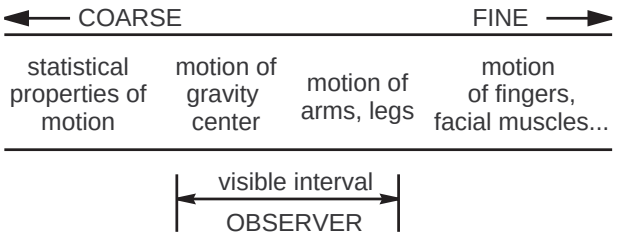

Figure 1: Human motion scale as seen from computer vision perspective.

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.

Computer vision based human motion 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 roughly determine the interval of scale that is visible to them.

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 [7]. 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. Example is provided in Figure 2.

Diagrams, shown in Figure 2 show the velocity and acceleration data, provided by two different systems when measuring the motion of the body of a human, following the triangular trajectory. The biomechanical motion analysis system (APAS, manufactured by Ariel Dynamics, Inc.) is at

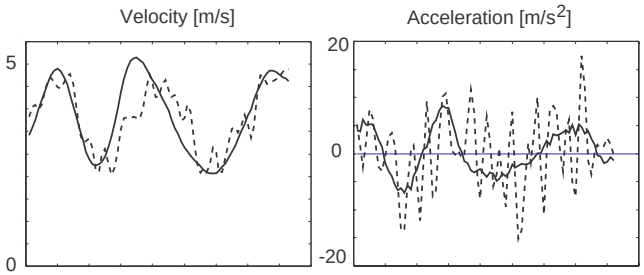

Figure 2: Output of two different tracking systems, measuring the same motion of person, following the triangular trajectory. Dashed line represents the output of the very accurate biomechanical motion analysis system, while the solid one represents the output of a simple blob tracker.

least ten times more accurate than the other system, which is simple blob tracker. Both systems provide the data about the motion of the human gravity center, projected on the ground plane.

Let us consider the problem of reconstructing the activity of the human, which has been measured by these two systems. The output of the second system (solid line) is much easier to interpret. It is very clear that it corresponds to the run along the triangular trajectory. Three major accelerations and deccelerations are clearly visible. On the other hand, high accuracy of the biomechanical motion analysis system only introduces irrelevant information into the big picture - the velocity and acceleration graphs include many subtle motion details. Individual steps of the person running are clearly visible.

Clearly, this data can be interpreted as three intervals of straight motion with three rapid turns, or, as number of fast accelerations and deccelerations quickly following one another. An important lesson can be drawn from this: the data with high level of detail is not always better - the choice of the right scale depends on the particular application.

In action recognition, the issue of the right scale becomes even 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 [8]. Sometimes, algorithms explicitly normalize video sequences around the perceived body center to suppress whole-body motion [6, 9] 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.

The right scale for human action recognition should be defined as the scale that contains useful information for particular action recognition problem. Furthermore, it may be possible that simultaneous observation of human motion on several clearly defined scales improves the recognition performance. For many action recognition problems we simply do not know which scales carry essential and usable information. Therefore, for the recognition system to work in practice, it should learn the desired scale of motion in some automatic way.

3 Experimental setup

Our experimental 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. Large amount of previous research makes sport an ideal test ground for human motion related experiments.

Each squash match is divided into sets, which may last from a few minutes up to 20 minutes or even more. In cooperation with sport experts we obtained digitized data from several matches, 18 sets total. They have been digitized at 384×288 pixel resolution and 25 frames per second. The recordings have been made on two different tournaments with different players.

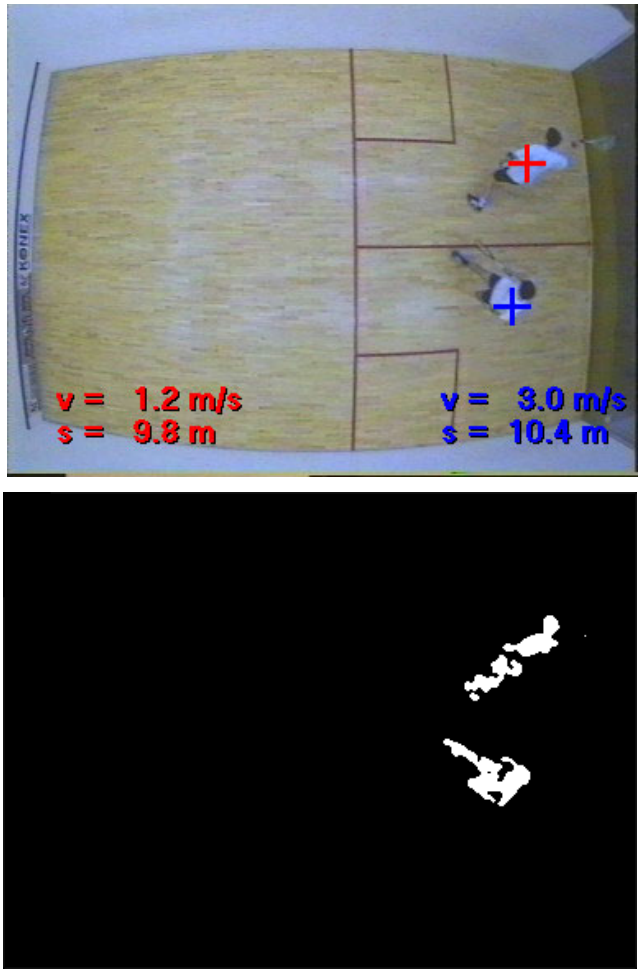

Figure 3: One frame of input video and a thresholded difference image.

3.1 Input data

Digitized video recordings of all sets have been processed to obtain trajectories of both players. The system was calibrated to provide position data in the court coordinate system. Position of players on every frame of the video was obtained by simple background subtraction tracking algorithm, which used the image of empty court as a reference. Figure 3 shows one of the frames from the video database and the corresponding thresholded difference image. Obtained player positions are marked with crosses. All tracking has been performed under supervision of a sport expert to prevent the tracker from switching the players.

3.2 Trajectory preprocessing

The sampling rate of obtained trajectories equals to the sampling rate of input video – 25 samples per second. To reduce noise in trajectories and increase the accuracy of velocity measurement, we smoothed the obtained trajectories using the Gaussian shaped kernel, which is shown in Figure 4. We processed x and y components of the trajectory separately, treating them as one-dimensional time-dependent signals

$$x'(t) = \frac{1}{2N_F+1} \sum_{i=-N_F}^{N_F} x(t+i) \cdot G(i), \quad (1)$$

$$y'(t) = \frac{1}{2N_F+1} \sum_{i=-N_F}^{N_F} y(t+i) \cdot G(i),$$

where $2N_F + 1$ denotes the width of the kernel, x' and y' are the smoothed components of the trajectory, and x and y are the components of the raw trajectory. G is the set of Gaussian coefficients which define the shape of the kernel. The precalculated set of $2N_F + 1$ coefficients in the range of Gaussian function $(-3\sigma, 3\sigma)$ was used. Kernel width $2N_F + 1$ is directly related to the intensity of the smoothing. Higher N_F yields smoother trajectories. After trajectory smoothing, velocity is obtained by differentiating the both components of the trajectory over time and calculating the length of the trajectory vector for each point of the trajectory. Field tests were conducted to evaluate tracker accuracy. Players were asked to follow different types of trajectories, marked on the court floor. In our case, we used filter width of 11 samples to smooth the trajectory data, and the accuracy of the tracker was found to be as follows:

- RMS position error: under 0.4 m
- RMS velocity error: under 0.6 m/s

The velocity data, obtained this way was the input data for our experiments, and in the remainder of the text we refer to it as *raw* or *unsmoothed* velocity, despite the trajectory smoothing.

4 The problem

The main assumption of our work is that certain properties of motion become evident only when observed at proper level of detail – *at the right scale*. To illustrate this, we used the velocity data from 18 sets of digitized video recordings, which corresponded to 18 sets of 6 different squash matches. Some information about each of the sets is shown in Table 1.

During the match, players run around the court, trying to hit the ball before it hits the ground for the second time. If one of the players fails to do this in time, match is stopped until the next serve. The interval between the error of one player and serve of the another is known as the ”passive

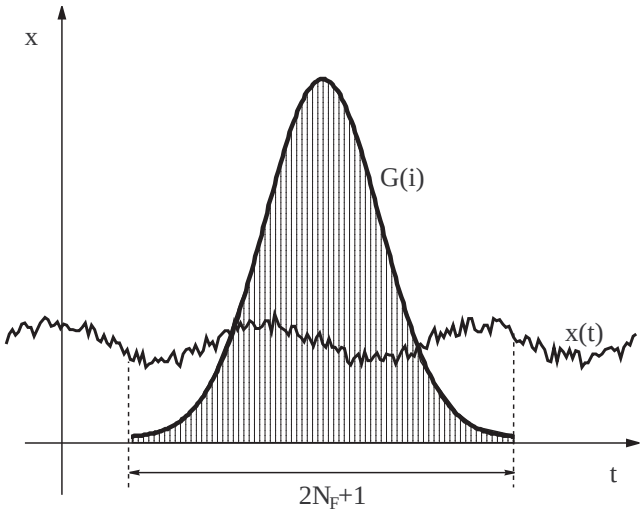

Figure 4: The shape of Gaussian kernel, used in trajectory and velocity smoothing.

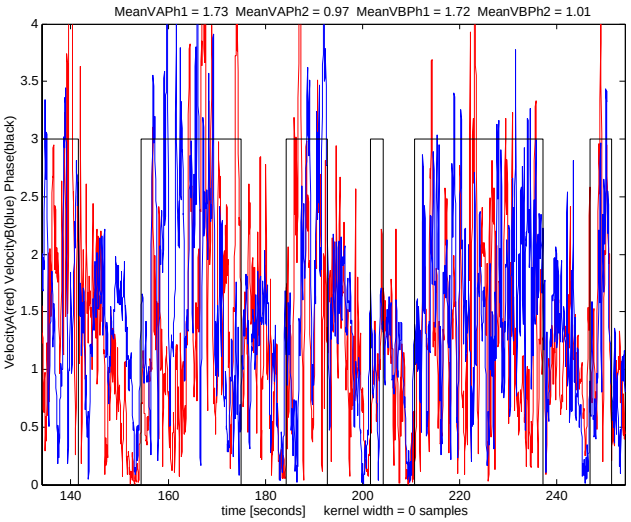

Figure 5: Raw velocities of both players and graph of active (high) and passive (low) phases for small segment of set number 1.

Set No.	Duration [s]	V_{active} [m/s]	$V_{passive}$ [m/s]	$T_{active}[s]$	$T_{passive}[s]$
1	931	1.72	0.99	548	383
2	261	1.64	0.96	154	107
3	406	1.59	1.03	274	133
4	985	1.65	0.91	646	339
5	893	1.68	0.84	553	341
6	1109	1.62	0.84	666	443
7	498	1.65	0.85	282	215
8	568	1.71	0.80	416	150
9	791	1.59	0.81	492	297
10	720	1.63	0.89	442	277
11	946	1.53	0.91	569	372
12	673	1.56	0.84	408	265
13	1101	1.57	0.89	638	458
14	756	1.46	0.83	426	330
15	586	1.61	0.78	336	250
16	530	1.57	0.81	261	269
17	523	1.70	0.79	254	269
18	593	1.74	0.96	379	214

Table 1: Set durations, average velocities of both players for active and passive phases and the total time spent in each of the phases. For this calculation, expert annotations were used to distinguish between the phases

phase" in sport science, and the time when players run and play is known as the "active phase".

Partition of the play to active and passive phases is important in match analysis, since the statistics (velocities, track length, duration) should be collected for each of the phases separately. The goal of our analysis was to automate the partitioning of match into active and passive phases. Sport experts provided manual annotation in terms of passive and active phases for each of the sets, for the whole duration of the sets. These annotations were used as a ground truth in our experiments.

4.1 The model of the play

A simple model of the squash play can be composed from the above information. The play is in the active phase when velocities of both players are high; the play is in the passive phase, when the velocities of players are low. This model may be also validated by observing the mean velocities in passive and active phases, shown in Table 1. Therefore, it should be possible to classify each sample of trajectory as being part of active (ω_1) or passive (ω_2) phase simply by arranging the velocities of both players into the two dimensional feature vector and feeding it into the classifier.

5 Experiments

If we look at the velocity graph, shown in Figure 5 of both players, it becomes clear that this simple model does not fit well into the obtained velocity data. Players do not always move slowly in the passive phase, and they are not always moving fast during the active phase. The motion in both phases is comprised of numerous accelerations and deccelerations, however it may be also observed that they *generally* move slower in the passive phase.

The two clusters of data are intermixed in feature space, as shown in Figure 6, and classification result is poor, with overall classification error of 32%. We used simple linear classifier [12] with the following setup:

- Method: Linear Least Squares
- Error estimation method: Holdout
- Percentage of training vectors: 2%
- No. of redraws (to reduce the effect of training vector sampling): 5

It is obvious that using raw data we cannot see the general principle of motion, on which our simple model is based. It should be emphasized that this is not due to *measurement errors* or *trajectory noise*, since these have been dealt with in the process of trajectory smoothing, before velocity calculation. The problem lies in the *scale of observation* - our velocity data contains details (motion of extremities, fast, short and sudden accelerations and stops) which obscure the *general impression* how fast the players move during the each phase. Therefore, this data has to be processed to remove the unnecessary details.

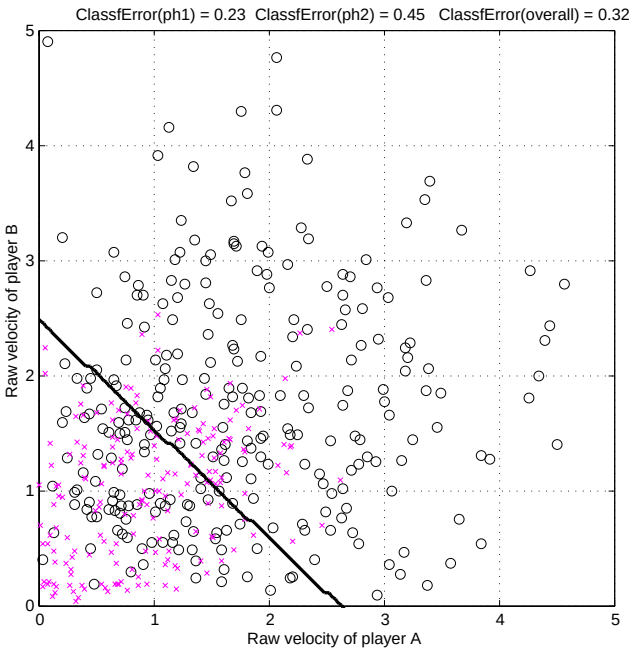

Figure 6: Feature space for raw velocity data and classification result for set number 1. Active phase samples (ω_1) are represented by circles, and passive phase samples (ω_2) are represented by crosses. To preserve graph clearness, only 2% of samples are shown.

5.1 Removing the details

We may remove subtle and unnecessary details from the motion data by *additionally* smoothing velocity data. The procedure and notation follows the description of trajectory smoothing, as specified in Equation 1, only this time the raw velocity data is smoothed, using the kernel of particular width.

To establish the right observation scale – the right amount of smoothing, which is controlled by kernel width $2N_F + 1$, we performed the number of experiments on each and every set from our database. The smoothing kernel width was varied from 0 samples (no smoothing) to 1000 samples in steps of 20 samples. Note that the sampling rate of both trajectory and velocity data is 25 samples per second. The classification was performed in each step, and classification error was observed.

6 Results

Figure 7 shows the dependence between the classification error and smoothing kernel width for one of the sets from our database. With increased smoothing the classification error is significantly lowered. However, from the particular kernel width on, error increases again.

Best filter width, which is related to *best scale of observation* may be read from those results. Distribution of such samples in feature space is shown in Figure 8. It is clear that in this case samples form only partially overlapped clusters, which decreased classification error by half in comparison to the samples of raw velocity.

Graph of smoothed velocities is shown in Figure 9. Such data clearly shows the intervals of low and high velocity,

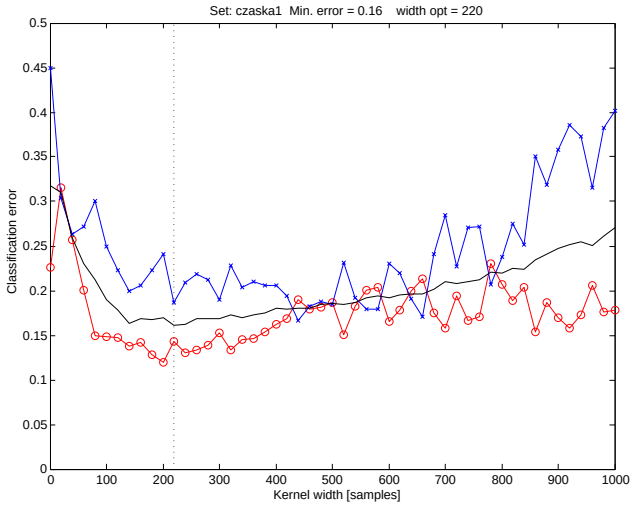

Figure 7: Relation between smoothing kernel width $2N_f + 1$ and classification errors for set number 1. Error for active phase samples (ω_1) is marked with circles, and error for passive phase samples (ω_2) is marked with crosses. Overall error is represented by the solid line. Vertical line marks the smoothing kernel width which produces the smallest overall classification error.

which roughly correspond to the passive and active phases of the squash play.

Experiments were performed on all 18 sets of data, and the results are shown in Table 2.

Set No.	Error (raw)	Error (smoothed)	Opt. kernel width [samples]	Average phase duration [samples]
1	0.32	0.16	220	306
2	0.29	0.16	320	242
3	0.32	0.18	300	254
4	0.25	0.12	180	362
5	0.24	0.13	160	328
6	0.26	0.13	140	347
7	0.29	0.17	120	270
8	0.17	0.10	80	406
9	0.25	0.12	260	271
10	0.25	0.12	340	346
11	0.27	0.16	300	364
12	0.27	0.16	100	336
13	0.30	0.15	200	278
14	0.38	0.19	360	350
15	0.26	0.14	220	293
16	0.28	0.17	200	288
17	0.24	0.13	240	311
18	0.26	0.12	300	280

Table 2: Overall classification errors for raw and optimally smoothed velocity data, optimal smoothing kernel widths and average phase durations for each of the sets.

It is obvious that classification results are consistently better, when the right amount of smoothing is applied to velocity data. Relation between smoothing kernel width and classification errors is in all sets similar to the graph, shown in Figure 7. It can be seen that in most sets the optimal kernel width lies in the interval between 100 and 300 samples. In this interval the classification error does not vary significantly, as shown in Figure 7. We believe that the arbitrary choice of smoothing kernel width within this range would not significantly affect the classification error. Additionally,

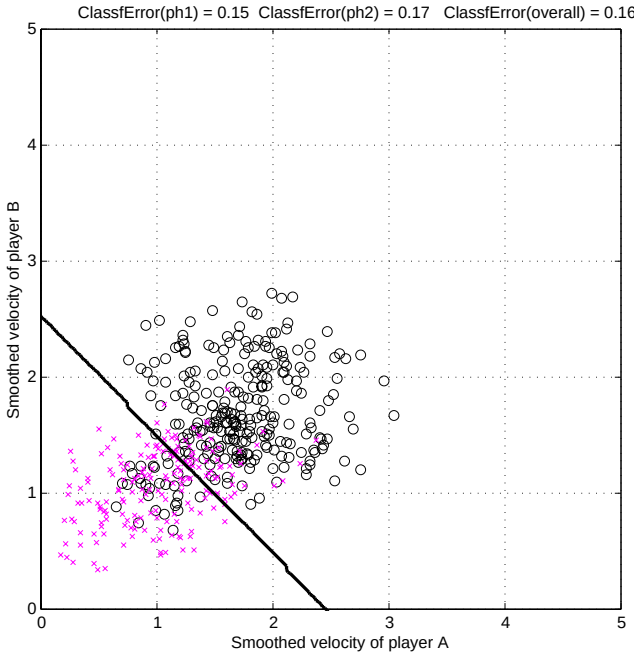

Figure 8: Feature space for optimally smoothed velocity data and classification result for set number 1. Active phase samples (ω_1) are represented by circles, and passive phase samples (ω_2) are represented by crosses. To preserve graph clearness, only 2% of samples are shown.

the optimal filter widths and average phase durations are of the same order of magnitude.

7 Conclusion

In this paper we addressed the problem of finding the right scale for a particular problem in human motion analysis. Our results show that the motion data with higher level of detail is not universally better, since it may obscure the important information with large amount of irrelevant data. We have shown that in our example, which is an activity recognition task in the sport play, the desired scale remains approximately the same, even if motion of different persons is observed at different times and in slightly different circumstances. This underscores our conclusion that the choice of right scale in human motion analysis remains primarily tied to the nature of particular human motion analysis problem and can be formulated with the question *"What do we want to see?"*.

References

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. In Dimitris. Metaxas and Irfan. Essa, editors, *IEEE Nonrigid and Articulated Motion Workshop*, pages 90–102. IEEE Computer Society Press, 1997.

[2] A. Ali and M. Farrally. A computer-video aided time motion analysis technique for match analysis. *The Journal of Sports Medicine and Physical Fitness*, 31(1):82–88, March 1991.

[3] B. A. Boghossian and S. A. Velastin. Image processing system for pedestrian monitoring using neural classification of normal motion patterns.

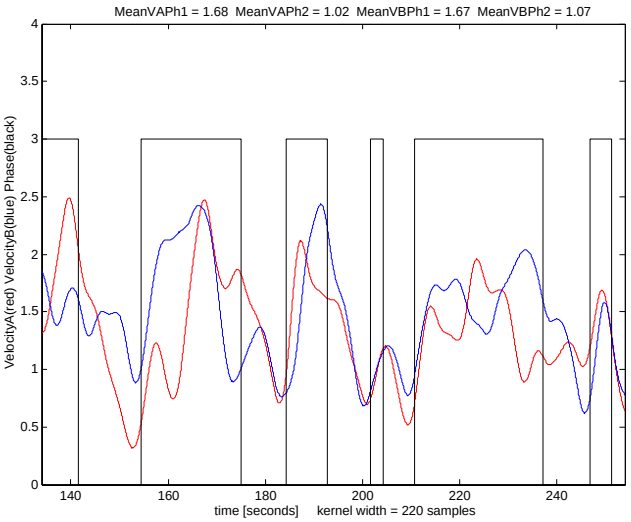

Figure 9: Optimally smoothed velocities of both players and graph of active (high) and passive (low) phases for small segment of set number 1.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36:585–601, 2003.

[12] E. Yom-Tov. Classification toolbox, december 2002. `http://tiger.technion.ac.il/~eladyt/Classification_toolbox.html`.

[13] L. Zeinik-Manor and M.. Irani. Event-based analysis of video. In *CVPR 2001*, pages II:123–130, Kauai, Hawaii, December 9-14 2001.