Multi-Scale Action Recognition in Squash Match

Janez Perš, Stanislav Kovačič
Faculty of Electrical Engineering
University of Ljubljana
Tržaška 25, 1000 Ljubljana, Slovenia
{janez.pers},{stanislav.kovacic}@fe.uni-lj.si

Abstract

Algorithms for human action recognition usually observe human motion only on particular level of detail. This approach requires complex algorithms to match the complexity of motion. High recognition rates are possible, when actions are distinct and clearly visible. However, this is not the case in many practical applications. To solve this we explore the possibility of developing more general action recognition algorithms by systematic reduction of complexity of human motion. The key to reducing complexity is in systematic decomposition of human motion to different scales, each representing different level of motion detail. 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, 2, 3], 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. Many researchers [4, 5, 6, 7, 8] 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 the particular problem.

In this paper, we explore the possibility of developing more general action recognition algorithms by systematic reduction in complexity of human motion. 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 is that human actions influence different scales of motion, and 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.

2 Scale-based human motion representation

Classification of video-based human motion analysis techniques is not uniform [1, 2, 3]. Nevertheless, the division to analysis of whole body motion vs. analysis of motion of the body parts is mainly undisputed. The first type of analysis 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. Scale-space representation of the world asserts that some properties of the observed object appear only when observed at a proper scale. However, human movement is a complex spatio-temporal phenomenon, 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.

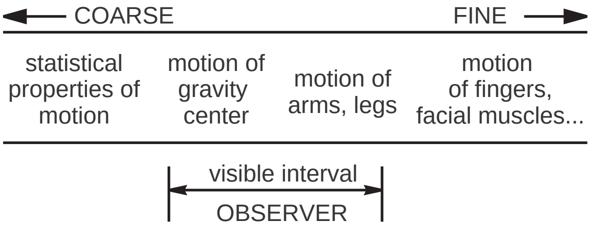


Figure 1: Human motion scale as seen from computer vision perspective.

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. Video-based human analysis algorithms essentially aim 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.

2.1 What is the right scale?

In the case of tracking, the observation scale is the main property of the tracking algorithm. If the scale is wrong for the particular application, algorithm will be labeled as inaccurate [9].

In action recognition, the issue of the right scale becomes more difficult. Most of the algorithms for action recognition focus on the single scale (movement of extremities, whole body motion, motion of facial muscles). However, 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.* 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.

Detecting hits in squash match represents a well defined real-world problem, as such annotations are needed for match analysis. Fig. 2 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, similar to [10] 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. 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.

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.

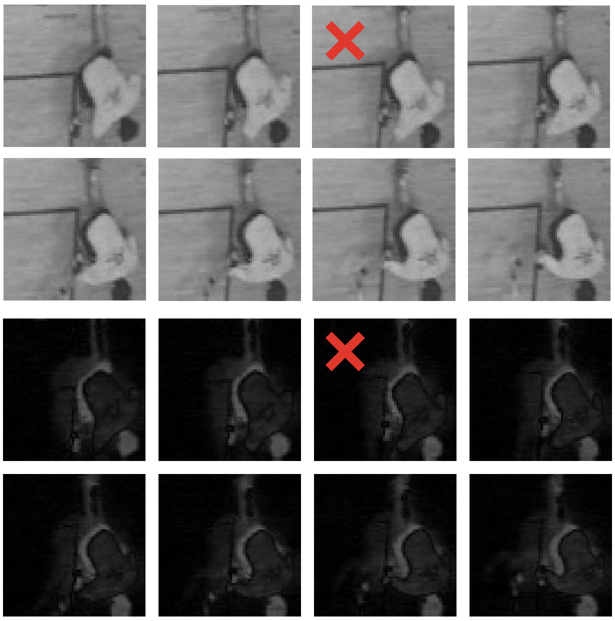


Figure 2: Image sequence of player, hitting a ball. Rows 1 and 2 show grayscale images, rows 3 and 4 show the images with background subtracted. Image, marked by X was annotated by expert as the exact moment of hit.

4.1 Training and classification

The training set S was split to two subsets, S_1 and S_2. S_1 (120 annotations) was used for training the classifier (LDA and PCA coefficients, decision boundaries), while S_2 (38 annotations) was used to estimate the preprocessing parameters which define the observation scale, by using exhaustive search in parameter space.

To classify the samples, we used Linear Discriminant Analysis (LDA) [11]. 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 of total classification rate on S_1.

4.2 Fine scale - the images

The motion of player arms carries significant amount of information about his actions. Due to low video resolution, appearance based method was used to extract this information.

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. 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, and the size of window was the first scale parameter. They were 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. 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 [11]. This is known concept of "eigensequences", as described by [12].

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*, [13]). Our tests have shown that these capture very little temporal action, which is visible only in lower-valued eigenvectors. The solution [11] 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.

4.3 Middle scale - the trajectories

Player actions may also cause typical motion of his body center, which is provided by our tracking algorithm. Therefore, trajectory shortly before and after the action was observed for typical patterns.

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. This is the first scale parameter, and the width of observation window is the second. 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. 3. 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 70% for ω_1 and 75% for ω_2.

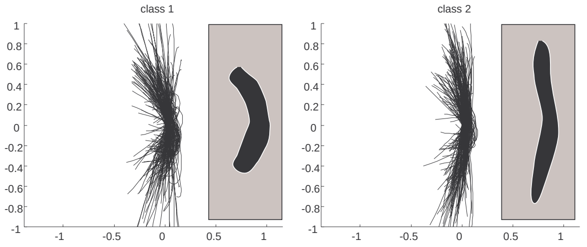


Figure 3: Normalized trajectories for the train set S.

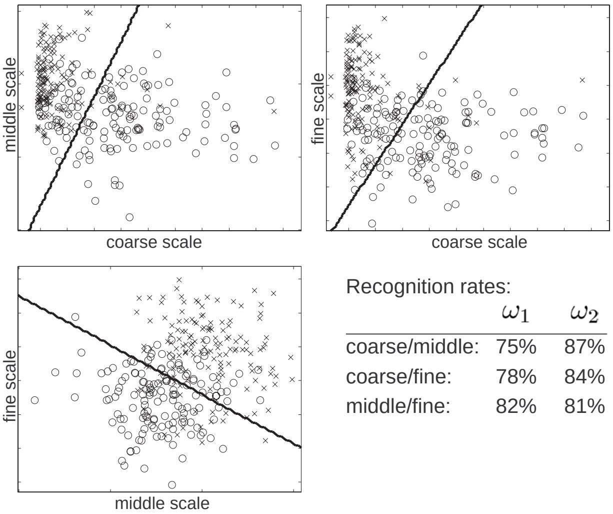
4.4 Coarse scale - statistical properties

Player motion exhibits certain statistical properties, when observed through long intervals of time. Due to the nature of the game, he is more likely to perform actions in certain areas of the court. This "apriori" probability distribution may be learned and used in action detection as well.

The player trajectory for the training set S was used to sample the probability of player presence at particular position, $P(pos)$. Joint probability, $P(hit, pos)$ was sampled from player trajectories at the moments of hits in S_1. Estimation of conditional probability of hit at particular position $P(hit/pos)$ was estimated using the formula $P(hit/pos) = P(hit, pos)/P(pos)$. Sampling was done by building 2D histogram, and smoothing it with 2D gaussian kernel. Kernel width (the only parameter) was established by observing its influence on classification of S_2. The samples from ω_1 and ω_2 were transformed to features simply by reading probabilities $P(hit/pos)$ at their spatial coordinates, and the decision threshold was learned from S_2. Using only this feature, the results of classification of the testing set T were 65% and 90% for ω_1 and ω_2, respectively.

4.5 Recognition on more scales

Our main assumption was that actions reflect themselves on different scales of human motion. To illustrate this, we joined the information from more scales. The test set T was transformed to the feature space, using algorithms for two scales at a time, 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. 4 shows the samples in this 2D space, for 3 different scale combinations.



Recognition rates:

	ω_1	ω_2
coarse/middle:	75%	87%
coarse/fine:	78%	84%
middle/fine:	82%	81%

Figure 4: Clusters and decision boundaries for 3 different combinations of scales. ω_1 - circles, ω_2 - crosses.

Diagonal placement of clusters show that the information from both scales works complementary and thus results in higher recognition rates. 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 for the three combination of scales are shown in Fig. 4.

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.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] J. Pers, M. Bon, S. Kovacic, M. Sibila, and B. Dezman. Observation and analysis of large-scale human motion. *Human Movement Science*, 21(2):295–311, 2002.

[10] J. Perš, G. Vučkovič, S. Kovačič, and B. Dežman. A low-cost real-time tracker of live sport events. In *Proceedings of ISPA 2001*, pages 362–365, Pula, Croatia, June 19-21 2001.

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

[12] 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.

[13] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, London, 2nd edition, 1990.