

---

# Human Activity and Pose Recognition using Space-time Correlation

---

**Hamed Pirsiavash**

Department of Computer Science  
University of California Irvine  
Irvine, CA 92617  
[hpirsiav@ics.uci.edu](mailto:hpirsiav@ics.uci.edu)

**Deva Ramanan**

Department of Computer Science  
University of California Irvine  
Irvine, CA 92617  
[dramanan@ics.uci.edu](mailto:dramanan@ics.uci.edu)

## Abstract

In this project we will implement and use space-time behavior based correlation algorithm to recognize the human activity and pose. Space-time correlation method measures the similarity between two distinct video clips based on the motion. We will use this algorithm in a nearest neighbor type classifier to find the best match to a new query in the database and then recognize the human activity based on prior known labels of the database.

## 1 Introduction

Human activity and pose recognition is a fundamental problem in many applications such as visual surveillance, video summarization, and video indexing [3]. In a very simple scenario, whenever we recognize the motion of human body and its pose, we can use it in indexing the vide database and then we may perform video search very efficiently based on the visual data. Also, in video summarization, usually, we are looking for video segments containing some particular types of motion or segments without them; hence, these algorithms can be very useful in these kinds of applications.

Some previous works use the model based approaches. They use constraints coming from kinematics and dynamics of the human body and using some models for body parts e.g., joints, and then estimate the parameters from video data. [4] Some other approaches use the visual data to estimate the human body motion and pose without any model [3]. Generally, they learn a function which maps high dimensional visual data to the motion space. These methods have good performance in initializing the model based approaches.

In this project, we will use a method closer to the second category. We will use space-time behavior based correlation method [1] to classify small sequences of human motion captured from different views. In section 2, the space time correlation method would be explained and in the next section, our classifier, implementation, and results would be discussed.

## 2 Space-time behavior based correlation

Shechtman et al have introduced space-time behavior based correlation as a novel method to measure the similarity between two video segments based on their motion [1]. They try to find the motion consistency between two space-time patches coming from two different video sequences. In some previous works, researchers have tried to calculate the optical flow for these two patches and then find the correlation between them. But these approaches have a lot of difficulties like aperture problem and high sensitivity of optical flow to noise. In this method, they try to define a similar correlation measurement avoiding explicitly calculation of the optical flow.

They choose small space-time patches e.g.,  $7*7*3$  patches from a video segment and calculate the gradient with respect to space and time and stack all of them in a matrix named  $G$ . We have

$$\nabla P_i \begin{bmatrix} u \\ v \\ w \end{bmatrix} = 0$$

$$\underbrace{\begin{bmatrix} P_{x_1} & P_{y_1} & P_{t_1} \\ P_{x_2} & P_{y_2} & P_{t_2} \\ \dots & \dots & \dots \\ P_{x_n} & P_{y_n} & P_{t_n} \end{bmatrix}}_{\mathbf{G}} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n \times 1}$$

Where  $(u,v,w)$  is the direction of pixels with equivalent intensity values and “ $n$ ” is the number of pixels in the patch. Figure 1 shows a patch and a  $(u,v,w)$  vector.

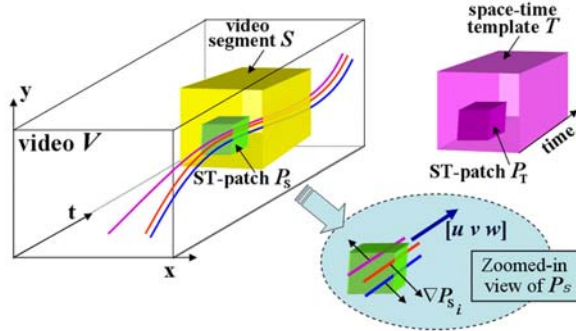


Figure 1. An ST patch and the vector containing pixels with equivalent intensity values.

We have

$$\mathbf{G}^T \mathbf{G} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3 \times 1}$$

$$\mathbf{M} = \mathbf{G}^T \mathbf{G} = \begin{bmatrix} \Sigma P_x^2 & \Sigma P_x P_y & \Sigma P_x P_t \\ \Sigma P_y P_x & \Sigma P_y^2 & \Sigma P_y P_t \\ \Sigma P_t P_x & \Sigma P_t P_y & \Sigma P_t^2 \end{bmatrix}$$

Hence, for a patch with consistent motion i.e, parallel equal intensity lines, M is a 3\*3 matrix for which the null space is the direction of equal intensity lines. Therefore, calculating the rank of M seems to be sufficient to investigate motion consistency in the patch. Note that in investigating the motion between two patches, we can concatenate the G matrices and calculate the resulting M as follows:

$$\mathbf{G}_{12} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \end{bmatrix}_{2n \times 3} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{2n \times 1}$$

$$\mathbf{M}_{12} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3 \times 1}$$

If the motions of both patches are consistent with each other, then this new matrix M would also have a null-space in the same direction.

However we know that if the spatial pattern of the patch is like an edge, it introduces a rank decrease in M and regardless of motion consistency of the patch, it would have a null space perpendicular to that special edge direction. In order to solve this degenerate case, they introduce the upper-right part of M as follows:

$$\mathbf{M}^\diamond = \begin{bmatrix} \Sigma P_x^2 & \Sigma P_x P_y \\ \Sigma P_y P_x & \Sigma P_y^2 \end{bmatrix}$$

Now, we have to calculate the rank increase by going from this 2\*2 matrix to our original 3\*3 matrix M. If this rank increase is 0, we have motion consistency and if it is 1, there are multiple motions in the patch.

They define rank increase using the eigen values of matrices as follows:

$$\Delta r = \frac{\lambda_2 \cdot \lambda_3}{\lambda_1^\diamond \cdot \lambda_2^\diamond}$$

And they introduce the inconsistency measure for each pairs of patches as follows:

$$m_{12} = \frac{\Delta r_{12}}{\min(\Delta r_1, \Delta r_2) + \epsilon}$$

Where the numerator is the rank increase for concatenated gradient vectors and “epsilon” is added to eliminate any possible division by zero. By summing the inverse of this measure over all possible patches for two space time windows from two video segments, we can find the similarity measure for those two windows. And by moving windows, we may find the correlation volume between two video sequences.

### 3 Implementation and results

In this project, we implemented the above mentioned method and used it in recognizing human motion in small video sequences. First, we tried the results of correlation algorithm on looking for a query containing only one type of motion (walking) in a video segment containing multiple motions. The videos are downloaded from [5] and after calculating the correlation, we have superimposed

the original video with a green ellipsoid on the peak value of the correlation. Figure 2 shows a sample frame of the output sequence. Input and output sequences can be downloaded from [http://ics.uci.edu/~hpirsiav/machineLearning/report\\_video1.zip](http://ics.uci.edu/~hpirsiav/machineLearning/report_video1.zip)



Figure 2. Result of the correlation algorithm on a walking query and a video segment containing multiple motions. Left to right: original video, query, superimposed result.

In using this method for activity and pose recognition, we used CMU MoBo database [2] which contains several sequences from 25 subjects walking in four different speeds and captured from six different views.

We chose six different motion classes from CMU MoBo database as follows:

1. Side-view fast walk (fastWalk/vr03\_7)
2. 45-degree-view fast walk (fastWalk/vr16\_7)
3. front view fast walk (fastWalk/vr07\_7)
4. Side-view slow walk (slowWalk/vr03\_7)
5. 45-degree-view slow walk (slowWalk/vr16\_7)
6. Front view slow walk (slowWalk/vr07\_7)

Some sample frames are shown in Figure 3. Then we chose one of the subjects randomly (04006) as the labeled dataset (train set) and recognized the motion class of the sequences of other 12 subjects (04002-04074) using nearest neighbor method. For each test sequence we calculate the space-time behavior based correlation volume and use its peak value as the similarity measure in NN classifier. The confusion matrix for this classifier is shown in Table 1. and the overall classification rate is %68.06.

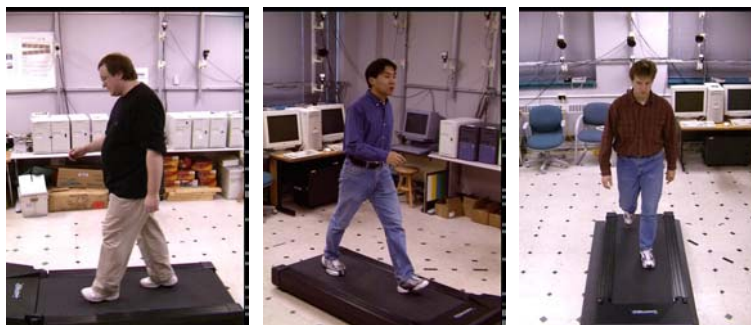


Figure 3. Some sample frames from different views. Left to right: fastWalk/vr03\_7, fastWalk/vr07\_7, fastWalk/vr16\_7

Table 1: Confusion matrix. Columns are test classes and rows are train classes.

Class Num	1	2	3	4	5	6
1. fastWalk/vr03_7	66.7	0	0	25.0	8.3	0
2. fastWalk/vr16_7	0	66.7	0	0	33.3	0
3. fastWalk/vr07_7	0	0	75.0	0	0	25.0
4. slowWalk/vr03_7	33.3	0	0	66.7	0	0
5. slowWalk/vr16_7	0	25.0	0	0	75.0	0
6. slowWalk/vr07_7	0	0	41.7	0	0	58.3

As shown in Table 1., most of the misclassifications come from making mistakes in capturing the speed of motion. By ignoring that part and combining fastwalk classes with the corresponding slowwalk classes, we would have three classes capturing only the view point. In this case, the confusion matrix shown in Table 2 would be much better.

Table 2: Confusion matrix by combining fastwalk and slow walk classes. Columns are test classes and rows are train classes.

Class Num	1	2	3
1. vr03_7	95.8	4.2	0
2. vr16_7	0	100	0
3. vr07_7	0	0	100

Space-time correlation method is computationally very expensive; hence, assuming a smooth surface (volume) for the correlation function, we implemented it hierarchically. This implementation is done in MATLAB platform and the running time for a pair of video sequences containing 30 frames each on a 3GHz Pentium 4 PC is about 3 minutes.

## References

- [1] Eli Shechtman and Michal Iran, (2007) Space-Time Behavior Based Correlation –OR– How to tell if two underlying motion fields are similar without computing them? *IEEE Trans on Pattern Analysis and Machine Intelligence (PAMI)* 29(11):2045-2056.
- [2] R.Gross and J. Shi (2001) *The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18*, Robotics Institute, Carnegie Mellon University.
- [3] Elgammal A., Lee C. (2004) Inferring 3D Body pose from Silhouettes using Activity Manifold Learning *IEEE CVPR*.
- [4] Fathi A., Mori (2007) G. Human Pose Estimation using Motion Exemplars *ICCV*
- [5] <http://www.wisdom.weizmann.ac.il/~vision/BehaviorCorrelation.html>