

# Configuration Estimation in Video

Alireza Fathi  
School of Computing Science  
Simon Fraser University, Burnaby, BC  
alirezaf@cs.sfu.ca

Greg Mori  
Vision and Media Lab  
School of Computing Science  
Simon Fraser University, Burnaby, BC  
mori@cs.sfu.ca

## Abstract

We introduce a behavior-based approach for finding body configuration and pose in the sequence of images in a video. A behavior-based similarity measure is employed to find the most similar motion not at a single frame but a sequence of frames between the query video and stored video templates. For each frame and frames adjacent to it, the most appropriate space-time matches are found from stored videos and manually marked joint positions are transferred to the matching place in the query. Joint positions are estimated by using probabilistic methods on the answers. Assuming that there is a stored template sufficiently similar in configuration and motion, the correspondence process will succeed. Using the behavioral-based similarity measure, the notion of 2-dimensional image correlation is extended into 3-dimensional space-time volume. This extension avoids false positives and negatives that exist in 2-dimensional correlation techniques. The presented approach is able to find joint positions accurately even in cluttered and moving background scenes.

## 1. Introduction

The problem we consider in this paper is to estimate the human body configuration in each frame of a motion video, not only by looking at individual frames but also by considering the action taking place at sequence of frames. This kind of approach will help to have a more precise estimation at end body limbs (e.g. hands and feet) compared to methods that just consider 2d images. This is because in this method motion fields are being matched and the end body limbs are the places with the fastest and clearest movements. If end body limbs are perfectly matched and the query person and the template one are nearly the same size, clearly we can assume that other parts of the body are matched perfect.

Action detection in space-time domain can be catego-

rized into two major classes. The first set of approaches use a 3D shape for representing the action in space-time. Pioneering work was done by Bobic *et al.* [2]. who use a motion history image to classify different actions. Blank *et al.* [1] use poisson equation to extract space time features such as local space-time saliency, action dynamics, shape structure and orientation. They show these features are useful for action recognition, detection and clustering.

The first class methods are robust in recognizing different actions but, they need background subtraction. This makes them useless at scenes with a moving background (e.g. at the beach or in a garden). The second class of approaches does not need any background subtraction. Dollar *et al.* [4] and Niebles *et al.* [10] use 3D interest points to describe different actions. Shechtman and Irani [9] use small space-time patches to find the best match for a space-time template in a longer video. Their method requires no background/foreground subtraction or segmentation, no prior training of activities, and no motion estimation or tracking. Their method works perfect at scenes with a moving and cluttered background and is able to find very complex behaviors in video sequences.

In this paper we have used the method in Shechtman and Irani [9] to match short space-time templates of stored videos with different frame sequences of our query video. For each frame in the query we get a list of best matches for upper, lower and the whole body windows. These results are used for ultimate pose estimation. We are using clustering and hidden Markov model to handle the mismatches between similar limbs.

The structure of the paper is as follows. In Section 2 we discuss related work. Our approach is described in more details in Section 3. In Section 4 the experimental results on CMU Mobo dataset [6] are presented. Section 5 concludes the paper.

## 2. Related Work

Until recently, most approaches to body configuration estimation are using silhouette images. Mori and Malik [7] use shape context to match individual body limbs and then try to assemble them to retrieve the pose. They get perfect results but their method can easily become confused because there can be many limb-like objects in the background. When occlusions are to be expected and background subtraction is not an option, for example because the camera is moving, Chamfer-based methods are among the most robust ones. Bobick and Davis [8] use both a hierarchy of templates and the Chamfer distance. This works excellent for difficult outdoor scenes. However, it seems to have a relatively high false detection rate.

Efros *et al.* [5] have taken motion into account. They estimate the optical flow at each frame by looking at its adjacent frames. Then they project it into a number of motion channels and use blurring to handle the noise. They use a nearest neighbor framework for recognition. Their method is so robust but becomes confused at scenes with cluttered background or textured clothing. This is because of aperture problem while estimating the optical flow which is interestingly handled in [9]. Dimitrijevic *et al.* [3] use Chamfer distance to match the whole body configuration in single frames of a space-time action. Then they use statistical relevance between all frames to fix the errors in 2D matches. This method is not considering the motion fields but it uses a sequence of frames for inference so it can be classified among the methods using space-time information for pose estimation.

## 3. Our Approach

Given a collection of stored template videos, each template video is tested to verify if it matches the query video in some place  $(x,y,t)$  in space-time domain. The method presented in Shechtmann and Irani [9] is used to search the query video by checking the global consistency between stored template with video segment centered around every space-time point. The global consistency between two such video segments is evaluated by computing and integrating local consistency measures between small space-time patches within the video segments. For each point in each video segment, a space-time patch centered around that point is compared against its corresponding space-time patch in the other segment. These local scores are then aggregated to provide a global correlation score for the entire template at that video location.

Different people with different clothes and different surrounding background, but with similar behavior, induce completely different space-time intensity patterns in a recorded video. To solve this problem, motion fields are used to correlate two space-time patches. space and time



Figure 1. Different Space-time Windows are used to match the query video.

gradients are calculated at each pixel inside a space-time patch  $(G)$ . Then  $G$  is multiplied by its transposed  $(G^T)$  to get the Gram matrix of  $G$  ( $M$ ) which is rank-deficient matrix. Rank increase is estimated by comparing  $M$  matrices of space-time patches with their addition. This helps to find the motion consistency and makes their method robust to different textures, colors and backgrounds.

In our approach we extract different space-time windows at some parts of the body from stored videos (Figure 1). Each space-time window is used to test if it matches at a space-time point of the query video. For each frame in the query video some good matches are found. The matched windows are used for labeling different body joints in the query frames. Always, there are some wrong matches. These matches are generated by similar motion fields of different body limbs. Sometimes the left leg of a person in query window is matched with the right leg of the one in stored video. Also sometimes hands have similar motions

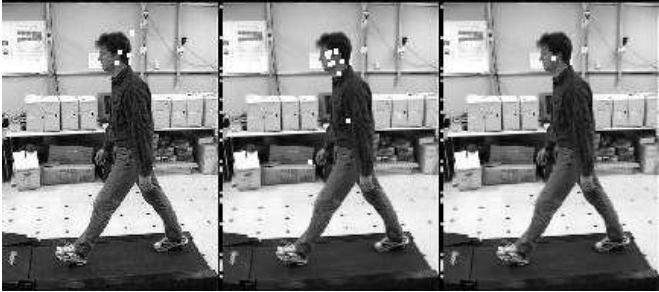


Figure 2. Different Space-time Windows are used to handle the mismatches between legs and arms. The first image shows the matches for head position from whole body windows. The second one contains the matches from upper body windows. These data are used to estimate the actual position of the head in the third image.

to the legs.

We have handled these problems as below. Some windows are used that contain a subset of limbs (*e.g.* upper body or lower body limbs). Also a window containing the whole body is used. This larger window is not perfectly matched but it is used to estimate where the person is located in the query video. We use the information gained by the whole body matches to eliminate false positives in smaller window matches. This will help us to ignore all wrong matches between hands and legs (Figure 2).

But Still there is a problem. Sometimes left and right limbs are mismatched (*e.g.* the right leg of the person in the query is matched with the left leg of the stored video or vice versa). To handle this problem we assumed that in most of the frames, the majority of matches are right ones. Having this assumption we used HMM to infer the right matches by looking at all frames. The details of HMM algorithm used on CMU Mobo database [6] is described in section 4. An example of applying HMM to find the position of left leg is shown in Figure 3.

## 4. Results

Experiments are performed on a same subset of images from the CMU Mobo database [6] as in Mori and Malik [7]. 9 subjects (numbers 04006-04071), 30 frames (frames numbered 101-130) are selected from the fast walk sequence for each subject. In all selected images the camera view is perpendicular to the direction of the subject’s walk (vr03\_7). The manually marked exemplar joint locations from Mori and Malik [7] are used.

We used this dataset to study the ability of our method to handle variations in body shape, clothing and backgrounds. However; the background is constant and we haven’t tried our method on datasets with moving background, but the results in [9] are taken as evidence that this method will

work in scenes with a moving background.

We have resized all images to 30% of the original size. Also images are converted to gray scale to make the matching more challenging. A set of nine experiments were conducted in which, each subject was used once as the query against a set of stored videos consisting of the image sequences of the remaining 8 subjects.

2D image positions of 14 keypoints (hands, elbows, shoulders, hips, knees, feet, head and waist) are estimated on the images by space-time behavioral matching of windowed video sequences. 3 space-time windows are used for each stored video: whole Body, Upper body limbs containing the head, shoulders, elbows and hands and the Lower body limbs containing the waist, hips, knees and feet. We have divided each stored space-time windowed video into 10 shorter space-time windows each of which consisting of 3 frames. As space-time matching process is computationally expensive we have used coarse to fine search for templates to enhance the speed. Then the query is searched precisely around the picks extracted from the coarse search. We have used  $14 \times 14 \times 3$  space-time patches in the matching process. The size of the patch represents the smaller cell that can be assumed to have a continuous motion. The best size for space-time patches is different in various scenes and has a significant effect on the strength and speed of matching.

As mentioned above, to handle the problem of mismatch between left and right limbs HMM is used. Our HMM algorithm has two states (left and right), each of which describing the position of left limb corresponding to the right one. There are 9 different observations each one describing the probability of what we are looking for being the left limb. This probability is computed by clustering the matches for the left limb and calculating the ratio of the left matches to all. The probability of transition from left or right state to the other one is  $\frac{1}{30}$ . This is because in a complete period of walking (30 frames) only it is the case in one frame that the left limb corresponding to the right one is at left and in the next frame will be at right. Some matching results after applying HMM are shown in Figure 3. The final joint estimation results are shown in Figure 4.

Our method is compared to Mori and Malik [7] in Table 1. To do this comparison, we have manually labeled left and right limbs. Automatically extracted results weren’t comparable because of large effective noises produced by mismatch between the similar limbs in a few frames.

## 5. Conclusion

Our primary contribution is a new joint estimation method using space-time behavior correlation instead of single silhouette estimates. We use some space-time windowed video segments to find the best matches for different



Figure 3. Hidden Markov Model is used to handle the mismatch between left and right limbs. Images in the first row contain the matches for left knee after eliminating noises. In the second row images HMM is used to find the actual position of the joint.

	5	6	10	11	13	14
<b>4011</b>	$16.1 \pm 9$	$17.7 \pm 12$	$32.8 \pm 10$	$34.4 \pm 26$	$20.4 \pm 8$	$23.4 \pm 11$
<b>4011</b>	$17.0 \pm 10$	$24.9 \pm 16$	$18.3 \pm 10$	$18.8 \pm 14$	$13.9 \pm 11$	$13.7 \pm 10$
<b>4013</b>	$26.9 \pm 13$	$27.6 \pm 12$	$16.7 \pm 9$	$17.1 \pm 9$	$18.7 \pm 10$	$17.6 \pm 10$
<b>4013</b>	$19.0 \pm 16$	$27.3 \pm 21$	$14.1 \pm 8$	$19.2 \pm 22$	$11.7 \pm 6$	$14.8 \pm 21$
<b>4068</b>	$26.7 \pm 16$	$42.2 \pm 40$	$21.6 \pm 12$	$25.3 \pm 18$	$12.9 \pm 8$	$21.3 \pm 12$
<b>4068</b>	$24.1 \pm 12$	$27.6 \pm 37$	$21.8 \pm 29$	$24.5 \pm 46$	$20.1 \pm 31$	$25.4 \pm 51$
<b>4070</b>	$25.2 \pm 8$	$32.5 \pm 12$	$19.8 \pm 8$	$23.8 \pm 12$	$19.4 \pm 16$	$34.2 \pm 15$
<b>4070</b>	$30.7 \pm 30$	$42.6 \pm 41$	$33.7 \pm 30$	$45.6 \pm 49$	$32.9 \pm 29$	$47.0 \pm 52$
<b>4071</b>	$29.5 \pm 16$	$22.7 \pm 12$	$20.2 \pm 12$	$24.0 \pm 12$	$24.1 \pm 19$	$30.5 \pm 17$
<b>4071</b>	$21.0 \pm 13$	$25.8 \pm 24$	$21.6 \pm 23$	$22.4 \pm 37$	$18.6 \pm 18$	$23.8 \pm 39$

Table 1. Mobo database sequence numbers versus joint position numbers. Each cell shows average error and standard deviation by our proposed method (top) and the Mori and Malik’s shape context [7] (bottom). Joint positions are: (5) Left Elbow, (6) Left Hand, (10) Right Knee, (11) Right Foot, (13) Left Knee, (14) Left Foot.

boy parts. We have estimated the joint positions from noisy data by using clustering and hidden Markov model. Currently our space-time correlation technique is not weighted. However, we can use Gaussian weights around the specific joints in different windows to find the exact positions of them. We have used a 1 dimensional HMM which can be extended to 28 dimensions to represent the position of different joints corresponding to each other on the plane. Our MATLAB program takes 1 minute to find the joint positions at each frame on a 1.7 GHz machine. Our future work is to implement this program in C. Also we will use more stored video sequences to be able to easily ignore the bad matching ones.

## References

- [1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402, 2005. 1
- [2] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. 1
- [3] M. Dimitrijevic, V. Lepetit, and P. Fua. Human body pose recognition using spatio-temporal templates. In *ICCV workshop on Modeling People and Human Interaction*, 2005. 2
- [4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*,



Figure 4. Results. Joint positions of the left shoulder, left knee, left hand and the joints on legs are shown.

2005. 1
- [5] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *Proc. 9th Int. Conf. Computer Vision*, volume 2, pages 726–733, 2003. 2
- [6] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical report, Robotics Institute, Carnegie Mellon University, 2001. 1, 3
- [7] G. Mori and J. Malik. Recovering 3d human body configurations using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(7):1052–1062, 2006. 2, 3, 4
- [8] C. F. Olson and D. P. Huttenlocher. Automatic target recognition by matching oriented edge pixels. *IEEE Transactions on Image Processing*, 6(1):103–113, 1997. 2
- [9] E. Shechtman and M. Irani. Space-time behavior based correlation. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 405–412, 2005. 1, 2, 3
- [10] Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(7):814–827, 2003. 1