

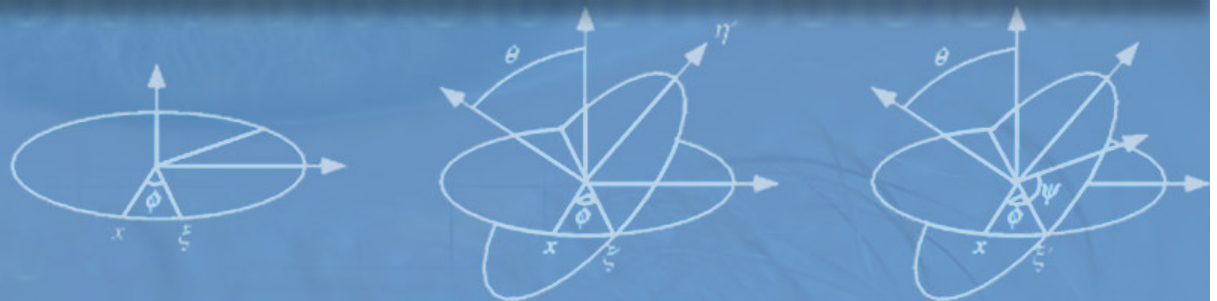


JHU vision lab

# Human Activity Recognition Using Multidimensional Indexing

By J. Ben-Arie, Z. Wang, P. Pandit, S. Rajaram, IEEE PAMI August 2002

H. Ertan Cetingul, 07/20/2006

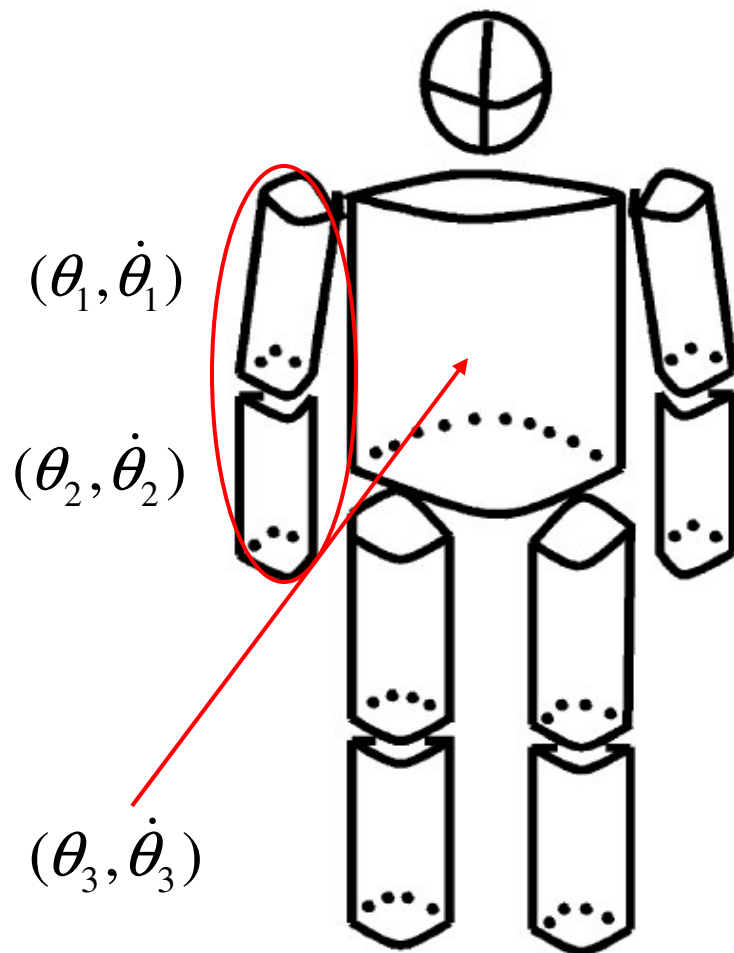


# Abstract

*“Human activity recognition from a sequence of angular poses and velocities of the main human body parts.”*

- An activity is represented by a set of pose and velocity vectors for the major body parts (hands, legs, and torso) and stored in a set of multidimensional hash tables.
- Each body part has a separate hash table which includes all the model activities.
- Recognize the activity invariant to the activity speed/time shift.
- It is claimed that it is robust to partial occlusion since each body part is indexed separately.
- It also has a view angle robustness of  $\pm 30$  degrees.

# The 9 body parts & feature vectors



5 tables: one 2D, four 4D

- The 9 body parts:
  - Torso (and head)*
  - Upper left arm*
  - Upper right arm*
  - Lower left arm (forearm + hand)*
  - Lower right arm*
  - Upper left leg (thigh)*
  - Upper right leg*
  - Lower left leg (calf + foot)*
  - Lower right leg*
- 18-dimensional feature vector
- $\theta_1$  :angle with x-axis
- $\dot{\theta}_1$  :angular velocity
- The angular velocities are calculated as the difference of angular positions of two successive frames.

# Theoretical Foundation (1)

- Human activity representation = concatenation of 18-dimensional subvectors (feature vectors)  $\mathbf{x}_i$  that describe the pose angles and angular velocities of 9 body parts.
- Pose angles are 2D projections of the actual 3D angles.
- Test vector  $\mathbf{Y}_t$  is compared with a set of model vectors  $\{\mathbf{Y}_m; m \in [1, M]\}$  where  $M$  is the number of the activity models in the database.
- The Maximum Likelihood Sequence Estimation (MLSE) is to determine the most likely sequence  $\mathbf{Y}_m$  given the observations  $\mathbf{Y}_t$ .
- Assumptions:
  - 1) Random differences between the subvectors  $\mathbf{x}_t$  and  $\mathbf{x}_m$  can be described as multivariate zero mean Gaussian distribution.
  - 2) Variations are conditionally independent from sample to sample.

$$\begin{aligned} P(\mathbf{Y}_t | \mathbf{Y}_m) &= P(\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, \dots, \mathbf{x}_{t_k} | \mathbf{x}_{m_1}, \mathbf{x}_{m_2}, \dots, \mathbf{x}_{m_k}) \\ &= \prod_{i=1}^k \frac{e^{-\frac{1}{2}(\mathbf{x}_{t_i} - \mathbf{x}_{m_i})^T C_x^{-1} (\mathbf{x}_{t_i} - \mathbf{x}_{m_i})}}{(2\pi)^{\frac{N}{2}} |C_x|^{\frac{1}{2}}}, \end{aligned} \quad (1)$$

# Theoretical Foundation (2)

- Likelihood function:

$$\begin{aligned} P(\mathbf{Y}_t | \mathbf{Y}_m) &= P(\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, \dots, \mathbf{x}_{t_k} | \mathbf{x}_{m_1}, \mathbf{x}_{m_2}, \dots, \mathbf{x}_{m_k}) \\ &= \prod_{i=1}^k \frac{e^{\left[ \frac{-1}{2} (\mathbf{x}_{t_i} - \mathbf{x}_{m_i})^T C_x^{-1} (\mathbf{x}_{t_i} - \mathbf{x}_{m_i}) \right]}}{(2\pi)^{\frac{N}{2}} |C_x|^{\frac{1}{2}}}, \end{aligned} \quad (1)$$

- Log-likelihood function:

$$\log P(\mathbf{Y}_t | \mathbf{Y}_m) = \sum_{i=1}^k \left[ \frac{-1}{2} (\mathbf{x}_{t_i} - \mathbf{x}_{m_i})^T C_x^{-1} (\mathbf{x}_{t_i} - \mathbf{x}_{m_i}) \right] - kG, \quad G = \log \left[ (2\pi)^{\frac{N}{2}} |C_x|^{\frac{1}{2}} \right].$$

- By Maximum-Likelihood:

$$\Omega = \arg \max_m \left( \sum_{i=1}^k \left[ \frac{-1}{2} (\mathbf{x}_{t_i} - \mathbf{x}_{m_i})^T C_x^{-1} (\mathbf{x}_{t_i} - \mathbf{x}_{m_i}) \right] \right).$$

# Indexing-based voting approach

- i. For each test vector  $\mathbf{x}_{t_i}$ , accumulate votes for all the models.
- ii. A model  $\mathbf{m}$  will accumulate an incremental vote of

$$\frac{-1}{2}(\mathbf{x}_{t_i} - \mathbf{x}_{m_i})^T C_x^{-1}(\mathbf{x}_{t_i} - \mathbf{x}_{m_i}) - G$$

for each test frame  $\mathbf{i}$ .

- iii. Repeat it for all the frames in the test sequence.
-

# Sampled poses/frames (K=?)

- Probability distribution  $p_{AB}(\bar{\delta})$  of the minimal multidimensional distance  $\delta$  between activities **A** and **B**.

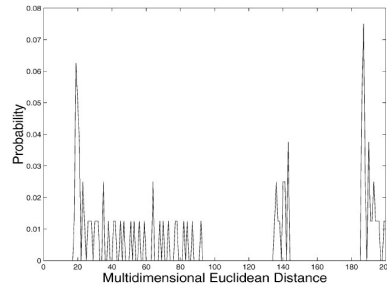


Fig. 2. Probability distribution of the multidimensional Euclidean distance for the jumping activity and the sitting activity. This distribution is quite sparse compared to the one in Fig. 3.

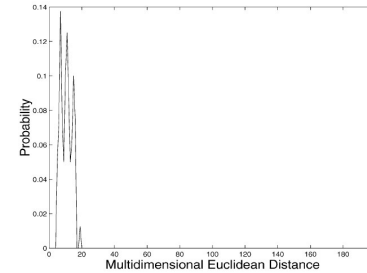


Fig. 3. Probability distribution of the multidimensional Euclidean distance between two different versions of jumping performed by different persons. Note that this distribution is quite concentrated in the low distance values compared with the one in Fig. 2.

- The area under this probability distribution curve below a threshold,  $P_{AB}(\delta \leq \gamma)$  indicates the probability that the test sequence is identified incorrectly (assuming that  $A \neq B$ ).

$$P_{AB}(\delta \leq \gamma) = \int_0^{\gamma} p_{AB}(\delta) d\delta = \alpha.$$

- The joint probability for a false matching in all the frames when using **K** test sampled poses can be obtained by,

$$P_{AB}(\delta_1 \leq \gamma \text{ and } \delta_2 \leq \gamma \cdots \delta_K \leq \gamma) \leq \alpha^K.$$

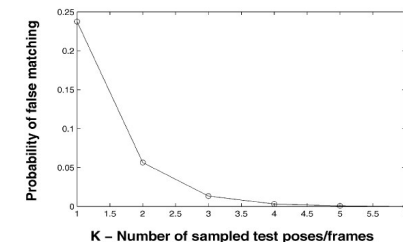


Fig. 4. Probability for false matching between jumping and kneeling as a function of  $K$ .

# Time shifts & speed variations

➔ *Embedded in the hash tables!*

- A priori unknown time shift  $\mathbf{a}_m$  (combining the votes with temporal correlation):

$$\Omega = \arg \max_m \left( \arg \max_{a_m} \left( \sum_{i=1}^k \left[ \frac{-1}{2} (\mathbf{x}_{t_i} - \mathbf{x}_{m_{i-a_m}})^T C_x^{-1} (\mathbf{x}_{t_i} - \mathbf{x}_{m_{i-a_m}}) \right] \right) \right),$$

---

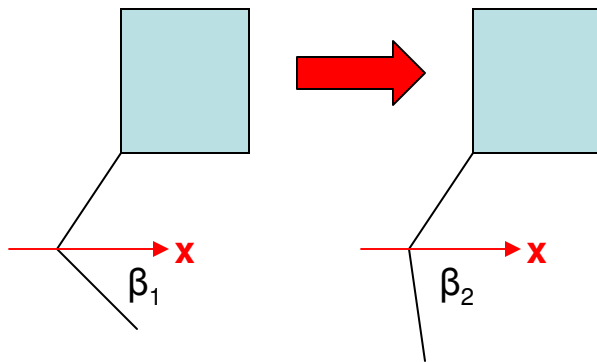
- Time scale  $\mathbf{s}$  + time shift  $\mathbf{a}_m$  (complex search for the optimum votes with various time scales and time shifts):

$$\Omega = \arg \max_m \left( \arg \max_s \left( \arg \max_{a_m} \left( \sum_{i=1}^k \left[ \frac{-1}{2} (\mathbf{x}_{t_i} - \mathbf{x}_{m_{s(i-a_m)}})^T C_x^{-1} (\mathbf{x}_{t_i} - \mathbf{x}_{m_{s(i-a_m)}}) \right] \right) \right) \right),$$

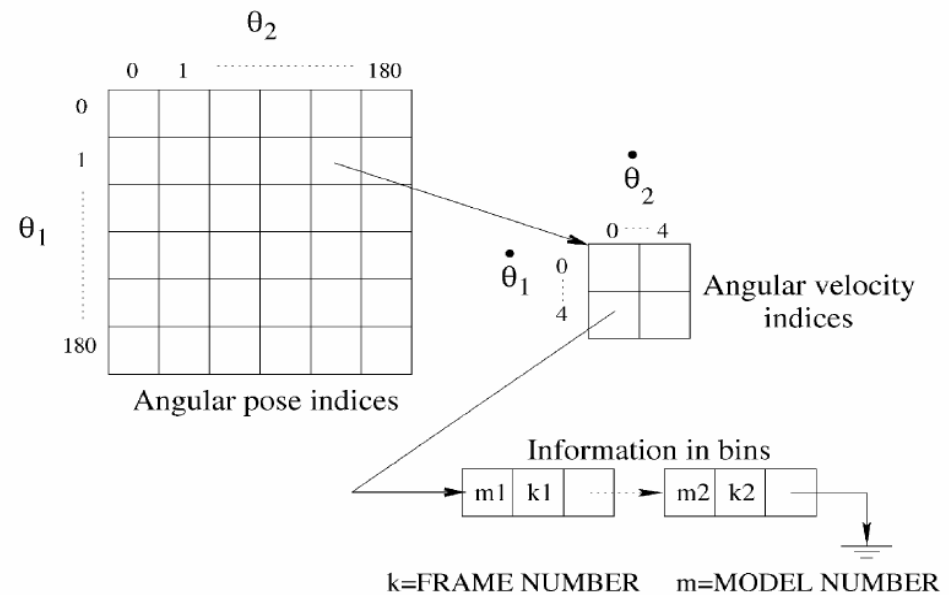


# Multidimensional indexing and voting

- The 2D Cartesian coordinates of all the major joints connecting the body parts are derived using a tracking procedure called EXM.
- Invariance to body size is achieved by transforming 2D coordinates to 2D angles.
- The hash table is 4D for the limbs and 2D for the torso.
- Every entry may include a set of different activity models which pertain to the same body part pose.



5 tables: one 2D, four 4D =  $\{h_1, \dots, h_5\}$



# Recognition scheme

- i. Voting for the individual body parts.
- ii. Combining the votes of the individual body parts for each test pose.
- iii. Final activity vote by integrating the votes of individual test frames.

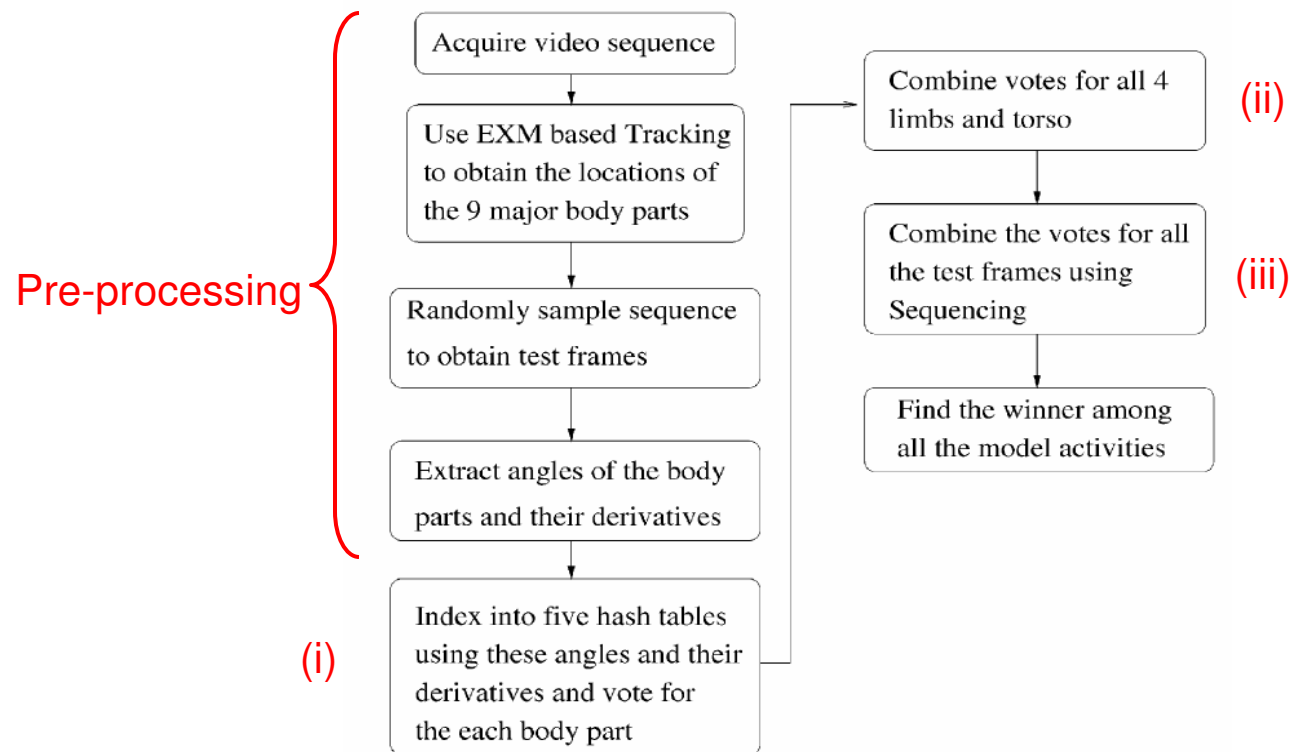


Fig. 7. The Flow Diagram of our recognition approach.

# Voting process (1)

- The voting scheme for each part  $h_i$  employs  $M$  1D arrays  $V_{mk}^{h_i}(t)$ ,  $m \in [1, M]$ , where each array corresponds to a different activity model and to  $k$ , which is the frame number of the test activity.
- To tolerate slight pose variations:  
Let  $b_i^k = (q_1^k, q_2^k, q_3^k, q_4^k)$  denote the quantized bin of one of the limbs ( $h_i, i \in [2, 5]$ ) for a test pose in test frame  $k$ .  
Let  $b'_i = (q'_1, q'_2, q'_3, q'_4)$  denote a neighboring bin in the corresponding hash table.
- Define the mapping function and start voting:

$$f(b, c, d, e) = \log e^{-\frac{1}{2} \left[ \left( \frac{b-b_0}{\sigma_b} \right)^2 + \left( \frac{c-c_0}{\sigma_c} \right)^2 + \left( \frac{d-d_0}{\sigma_d} \right)^2 + \left( \frac{e-e_0}{\sigma_e} \right)^2 \right]}$$

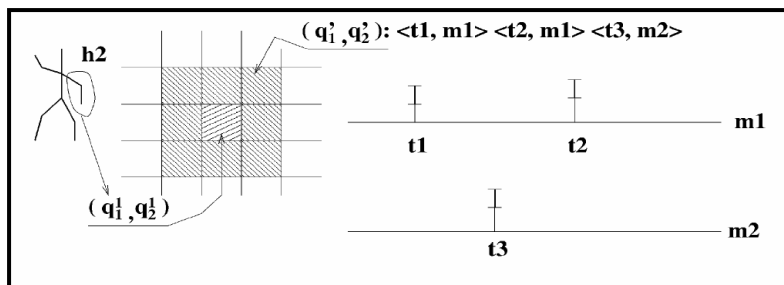


Fig. 8. A voting example of the left arm. On the left, the center square ( $q_1^1, q_2^1$ ) of the grid represents the bin from the pose of the left arm and the surrounding squares are neighboring bins. The upper-right bin ( $q_1^2, q_2^2$ ) contains three entries from models  $m1$  and  $m2$ . These votes are described by the bars on the right diagram. This diagram describes two 1D voting arrays for activity models  $m1$  and  $m2$ .

1<sup>st</sup> stage

$$V_{mk}^{h_i}(t) += (f(|q_1^k - q'_1|, |q_2^k - q'_2|, |q_3^k - q'_3|, |q_4^k - q'_4|))$$

$$V_{mk}^{h_1}(t) += f(|q_5^k - q'_5|, |q_6^k - q'_6|)$$

2<sup>nd</sup> stage  $\rightarrow V_{mk}(t) = \sum_{i=1}^5 V_{mk}^{h_i}(t).$

3<sup>rd</sup> stage  $\rightarrow V_m = \sum_{k=1}^K V_{mk}(L_k).$

# Voting process (2)

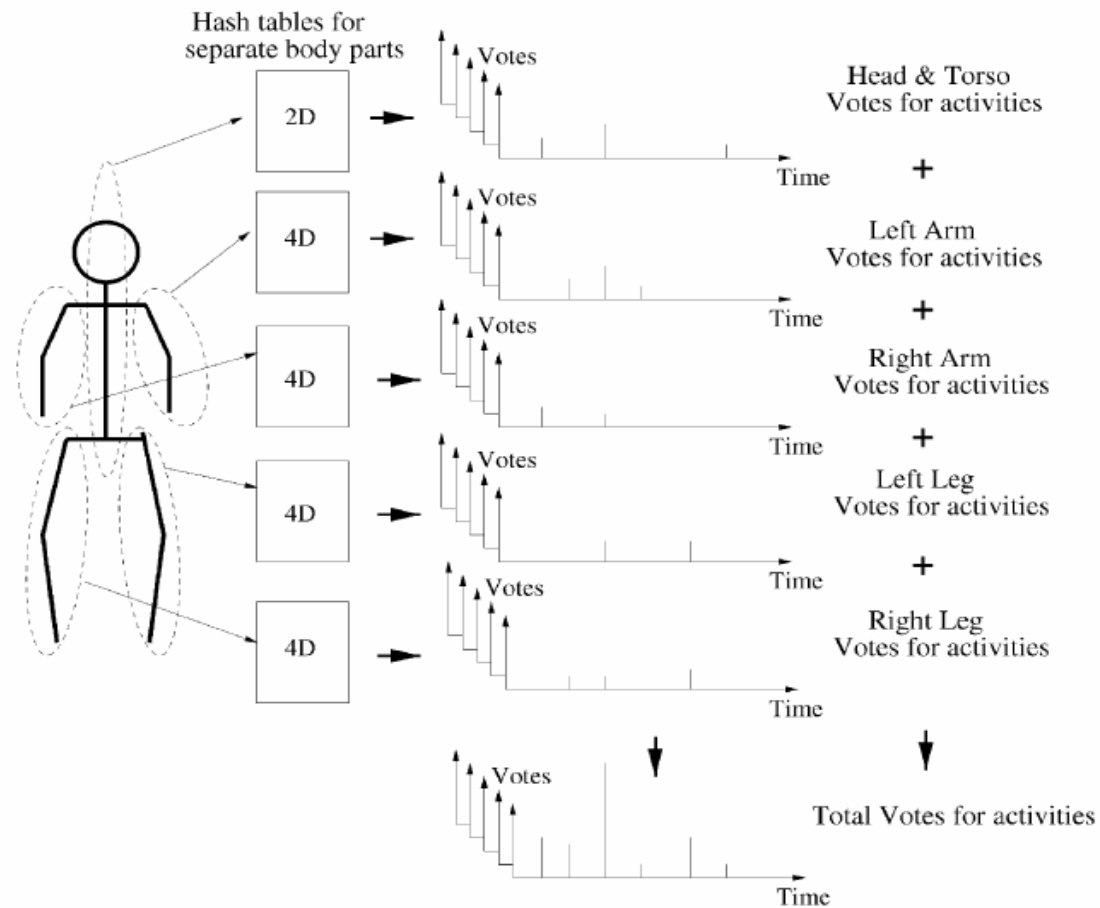


Fig. 9. A diagram of the whole voting process which illustrates how voting takes place for the different body parts for different model activities and the way in which the votes for different body parts are combined.

# Numerical Example

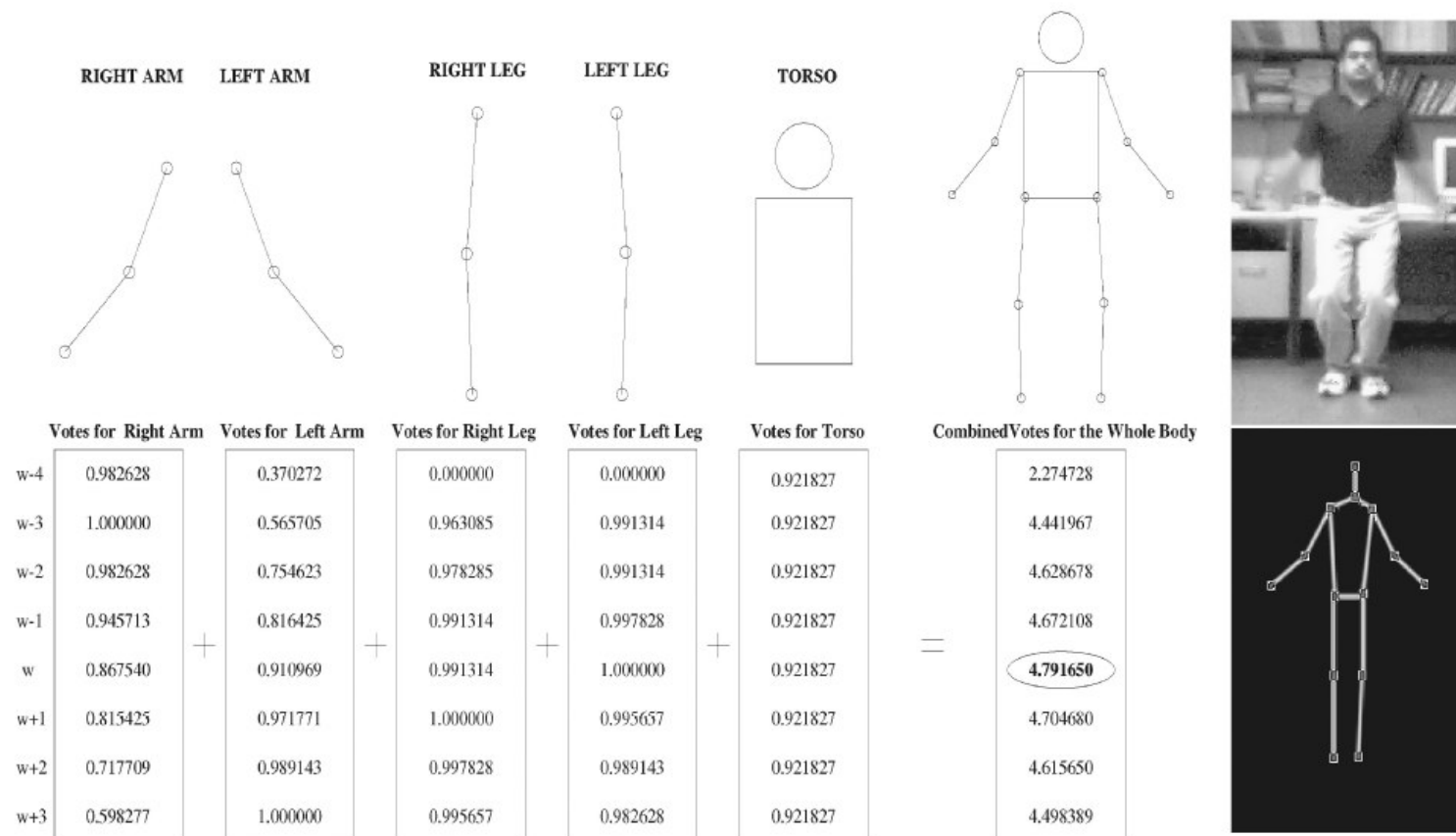





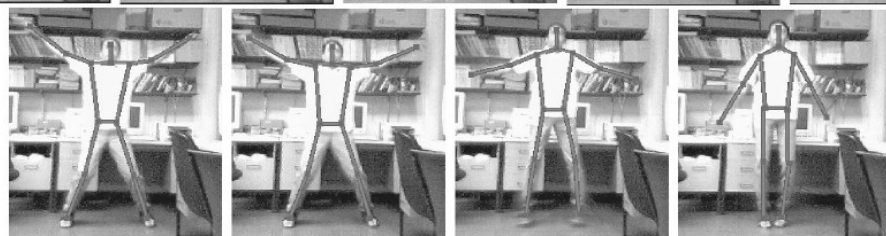
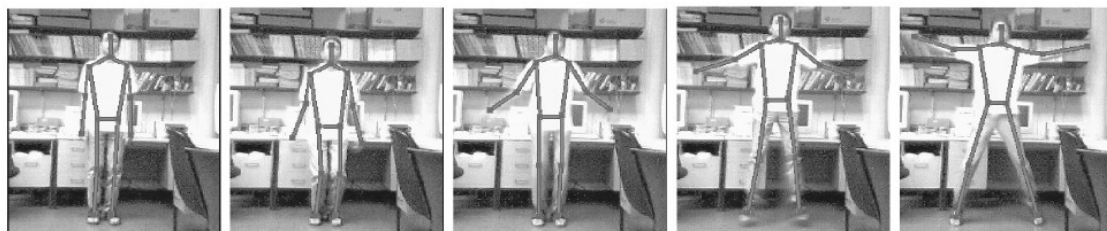
Fig. 10. A numerical example illustrating the process of voting for the five body parts for a pose corresponding to a jumping test sequence. The votes shown are with respect to different frames of model jumping activity stored in the hash table. The votes are shown for a subset of model frames which include the winning pose and its neighbors. The frame numbers of the poses which are shown are represented in terms of the winning frame  $w$ . The test pose which is being voted for is shown in the picture in the top right-hand corner and the model pose which received the highest vote is shown in the bottom right-hand corner.

# Clarification about t

- 3 activities (m1, m2, m3)
- 9 time instants (t1 to t9)
- 3 poses (k=1,...,3)
- Let the votes be:
  - k=1  m2t6 > m2t7 > m1t3 > m1t5 > m3t6 > m3t2
  - k=2  m2t5 > m1t2 > m2t7 > m3t3 > m1t4 > m3t4
  - k=3  m1t6 > m2t9 > m3t5 > m3t1 > m2t3 > m1t1
- you are allowed to add:  
m2t6 + m2t7 + m2t9 OR m1t3 + m1t4 + m1t6 OR m3t2 + m3t3 + m3t5
- And select maximum!

$$V_m = \sum_{k=1}^K V_{mk} (L_k).$$

# Sample frames



Jumping sequence



Sitting sequence

# Recognition Results

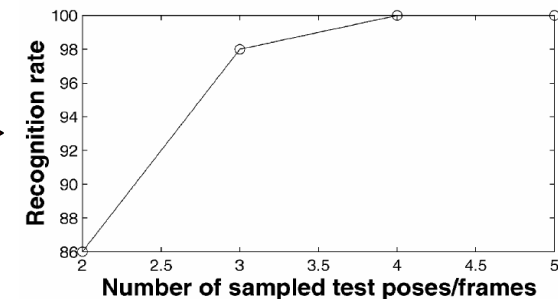
TABLE 1  
Average Votes of Activity Sequences for the Voting/Sequencing with Angular Pose and Velocity-Based Voting

	Jump	Kneel	Pick	Put	Run	Sit	Stand	Walk
Jump	<b>12.31</b>	3.91	1.97	2.00	2.18	2.00	1.20	3.55
Kneel	4.90	<b>9.99</b>	3.20	2.77	2.20	2.18	2.40	3.80
Pick	0.67	2.00	<b>8.00</b>	2.40	1.97	1.90	3.80	1.36
Put	1.95	2.58	3.10	<b>8.37</b>	1.58	4.71	2.50	1.74
Run	2.00	2.25	1.40	1.70	<b>3.23</b>	1.40	1.50	2.90
Sit	1.36	1.73	3.00	4.20	0.90	<b>8.60</b>	3.40	1.60
Stand	0.00	0.55	2.34	1.86	1.23	3.50	<b>9.90</b>	0.63
Walk	3.40	3.18	1.97	1.60	2.61	1.50	1.16	<b>5.75</b>

92 videos  
24 for training  
28 for view test  
40 for recog test

The rows correspond to test activity, while the columns correspond to the model activities. The best score in each row is in boldface numerals. The method yields correct recognition since the scores along the diagonal are the highest in each row.

100% recognition with 4 poses/frames →





# View-angle dependence

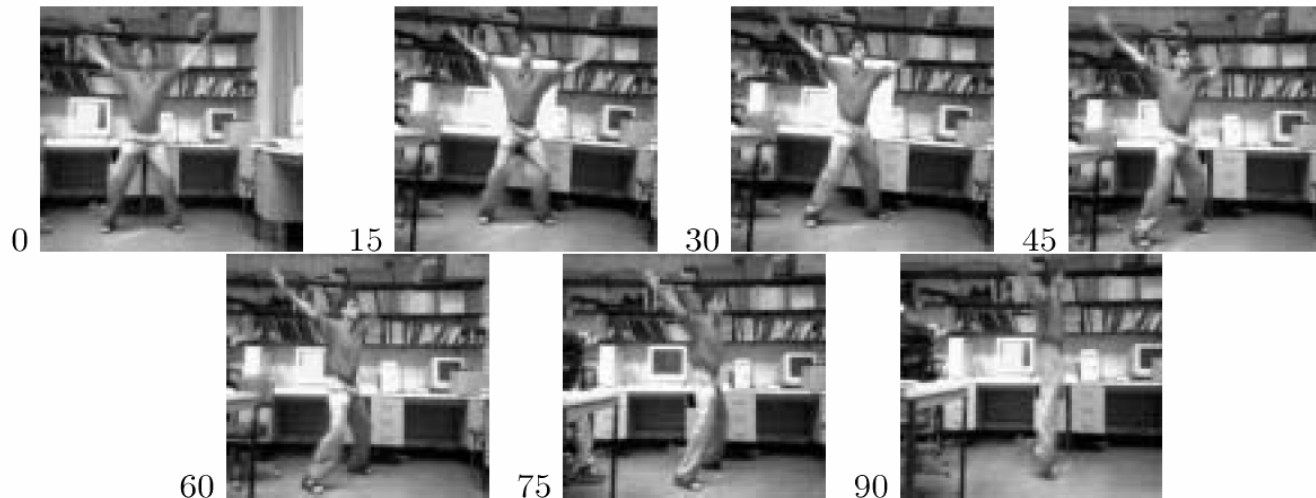
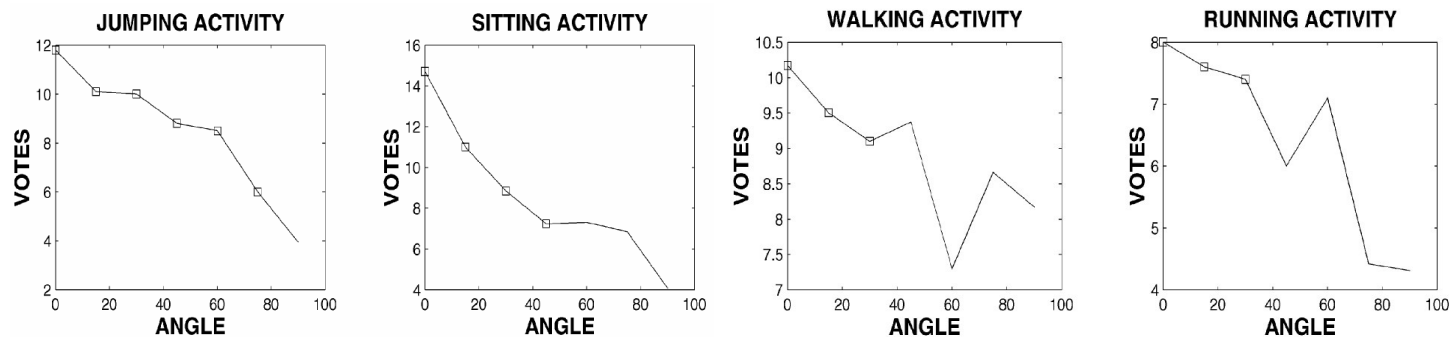


Figure 16: Frames of jumping sequence for different views of the activity. The number on the lower left corner of each of the images represent the azimuth angle (in degrees) at which the activity is viewed.



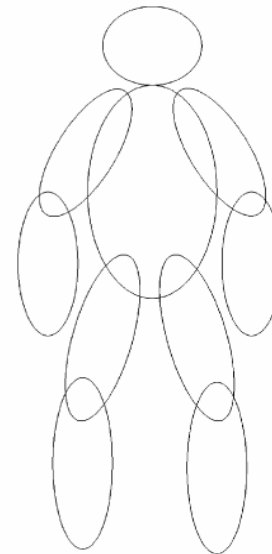
# Results with occlusion

TABLE 2  
Correct Recognition Rate under  
Occluded Conditions of Body Parts

Non-Occluded Body Parts	Correct Recognition Rate
1 Arm and Torso	70%
1 Leg and Torso	70%
1 Arm and 1 Leg	87.5%
1 Arm, 1 Leg and Torso	92.5%
Arms and Torso	85%
Legs and Torso	80%
Arms, 1 Leg and Torso	97.5%
Legs, 1 Arm and Torso	95%

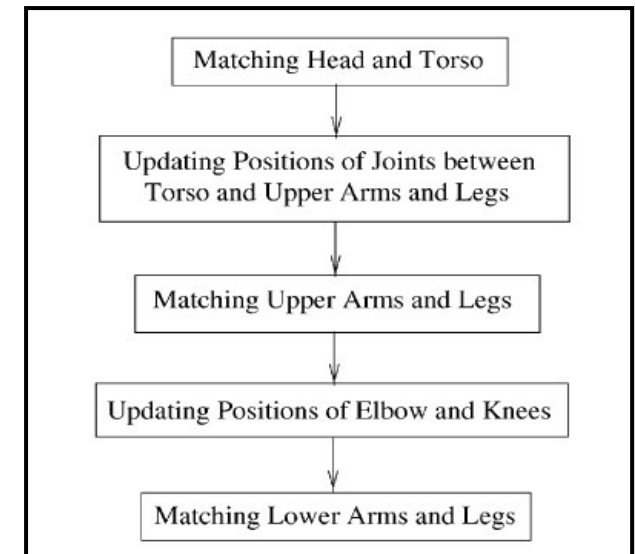
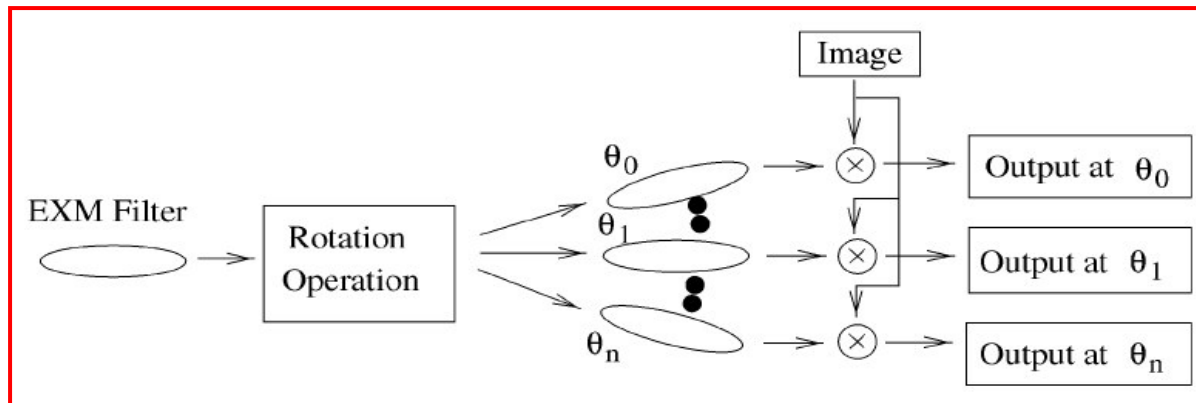
# Appendix : Expansion Matching (1)

- ➔ “The EXM filter is an efficient template matching approach and provides good results since it relies on image features with medium/high frequency content such as texture and edges.”
- *J. Ben-Arie and K.R. Rao, “Optimal Template Matching by Non-Orthogonal Image Expansion Using Restoration,” Int’l J. Machine Vision and Applications, vol. 7, no. 2, pp. 69-81, Mar. 1994.*
- *J. Ben-Arie and K.R. Rao, “A Novel Approach for Template Matching by Nonorthogonal Image Expansion,” IEEE Trans. Circuits and Systems for Video Technology, vol. 3, no. 1, Feb. 1993.*
- *Robust to partial occlusion.*
- *After representing each body part with ellipse, we construct an EXM filter for each body part, and the tracking is performed by application of these EXM filters successively and searching for maximum response.*



# Expansion Matching (2)

- The EXM is based on expanding the signal with respect to a set of basis functions that are all shifted versions of the template. In practice, the template, which serves as a basis function, is translated to all the candidate locations in the image. The magnitude of the expansion coefficients obtained at a particular location signifies the extent of the presence of the template at that location.
- EXM optimizes a novel Discriminative Signal-to-Noise Ratio (DSNR), which considers as unwanted clutter all the responses not at the center location, it achieves a very sharp output peak where the template matches the image.
- The tracking is performed by applying the set of 9 EXM filters represented in elliptical regions to each frame in the video stream.



# Expansion Matching (3)

- After a complete round of filtering and searching, the set of the part EXM filters need to be updated to accommodate the new positions, orientations, and lighting variations.

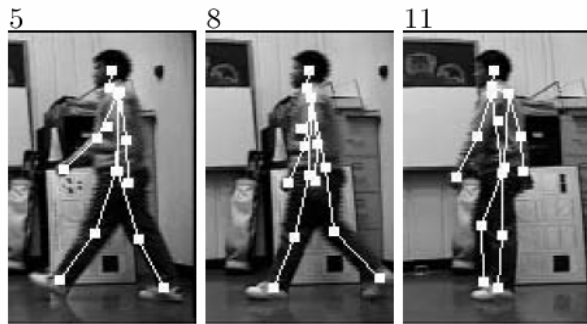


Fig. 14. The frames of a video stream of a man walking overlaid with wire frames representing detected parts. (a) Frame 5, (b) frame 8, and (c) frame 11.

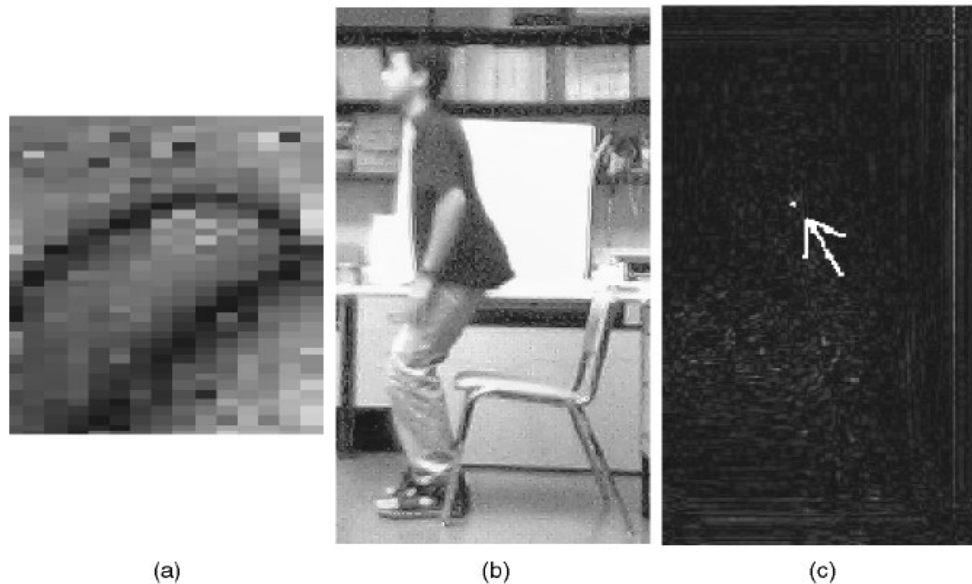


Fig. 21. An example of the detection of an elbow using EXM. (a) Shows the filter which corresponds to the elbow, (b) is a frame of the sitting sequence, and (c) is the result of application of the EXM filter (a) on the image in (b). Please note the strong peak found (marked by the arrow).