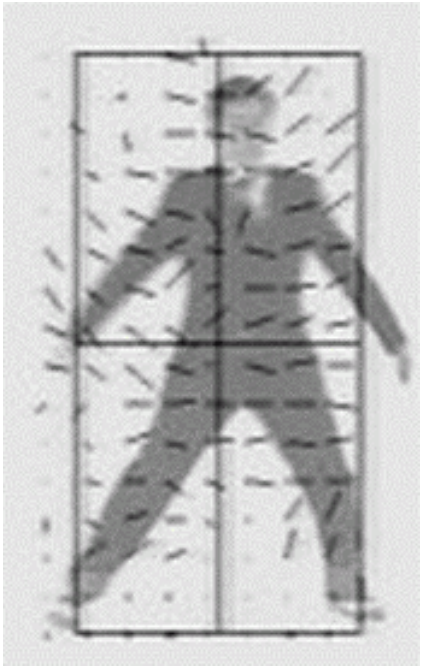
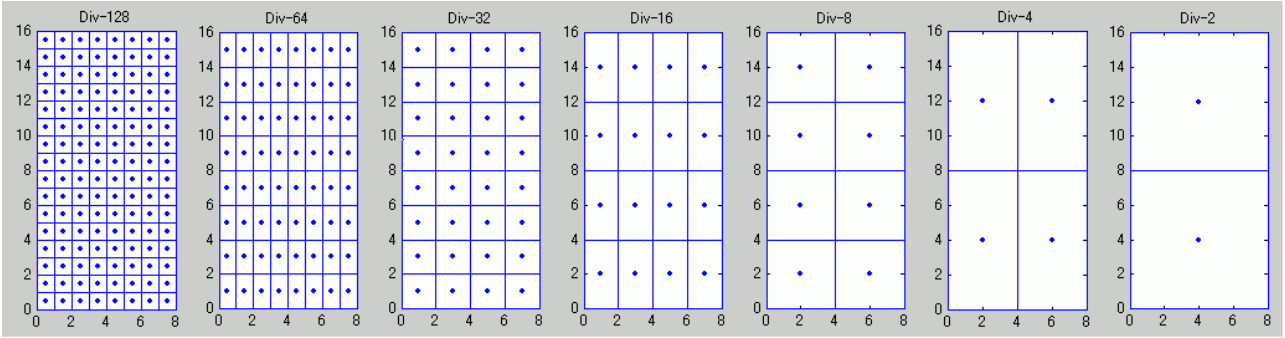
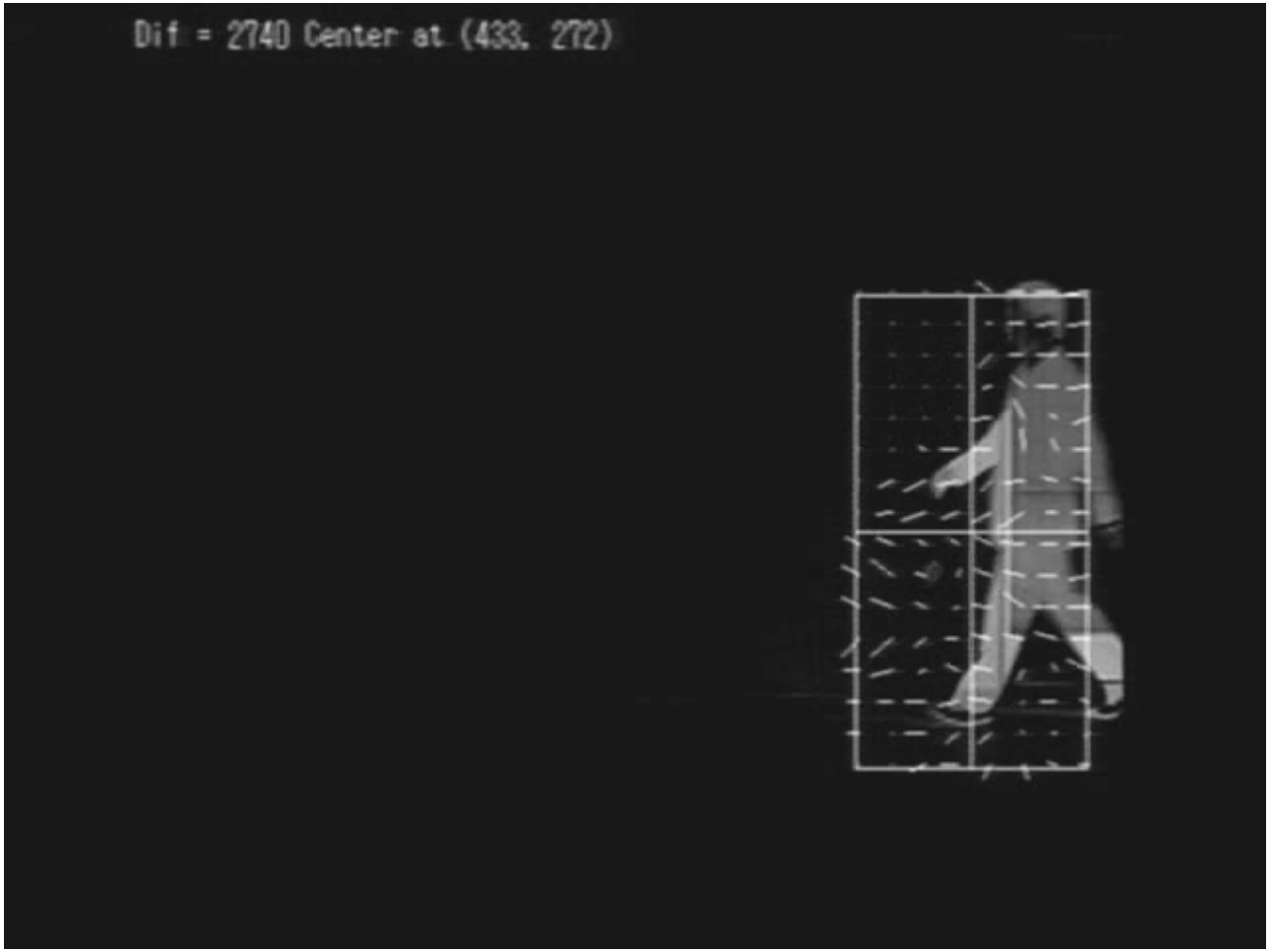
Recognizing Human Activities in Video by Multi-resolutional Optical Flows



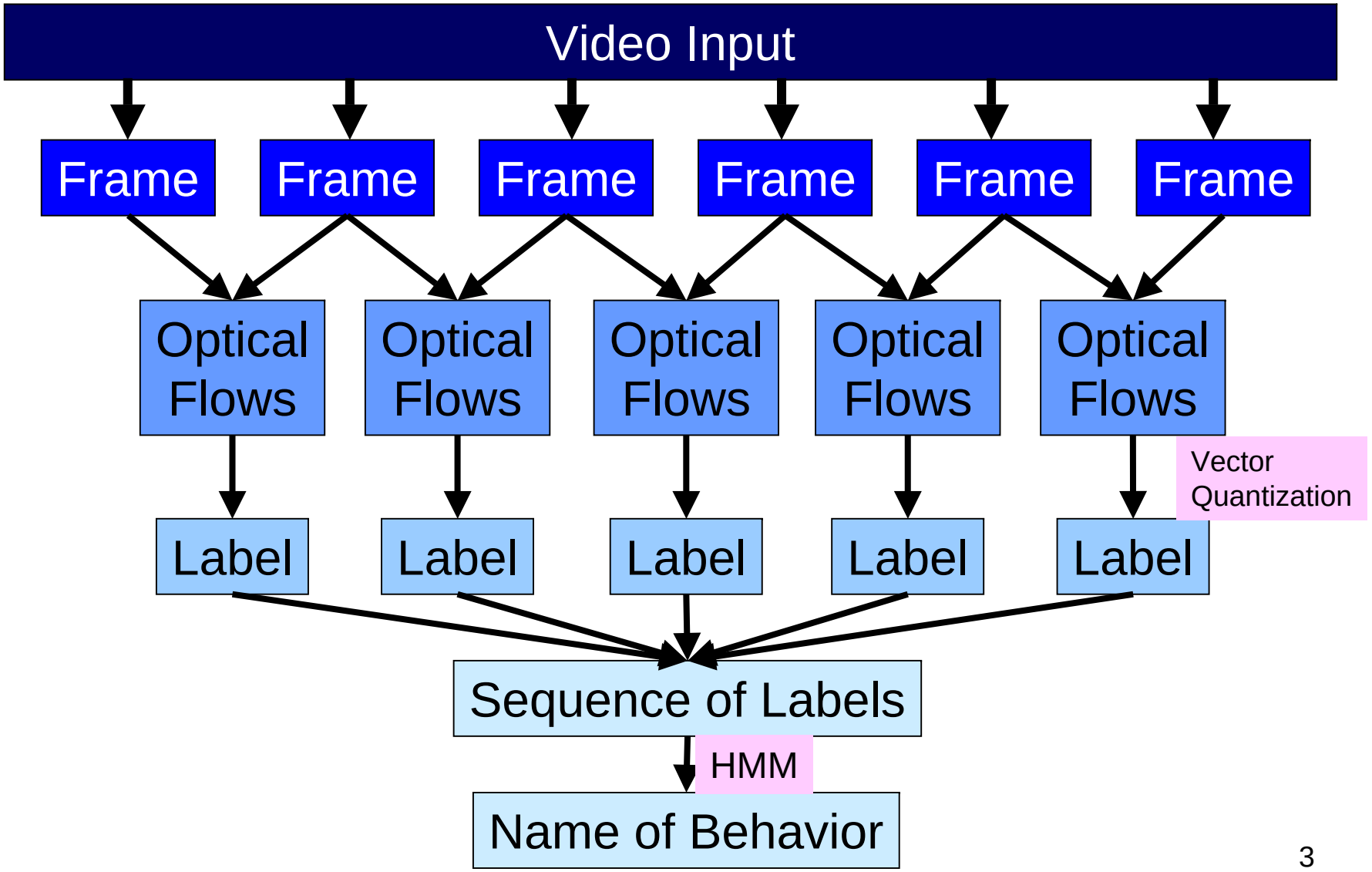


Toru Nakata
Digital Human Research Center,
AIST, Japan.

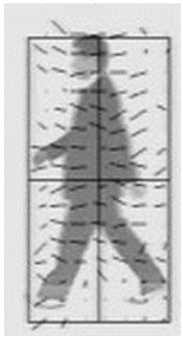
Recognize whole-body behaviors



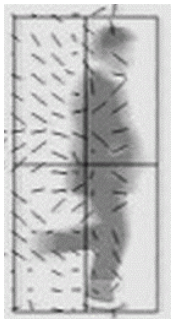
Overview of the process



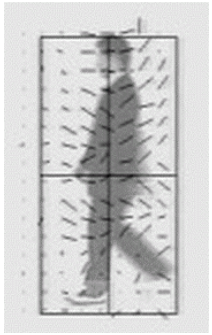
Behavior samples



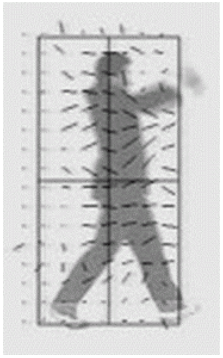
(A)



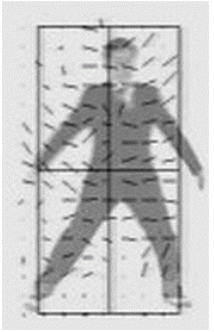
(B)



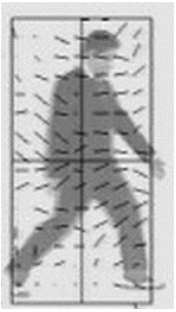
(C)



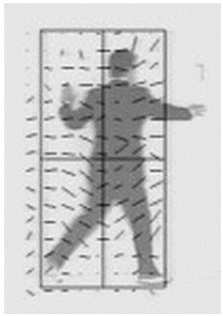
(D)



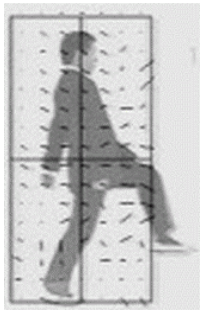
(E)



(F)



(G)



(H)

(J)

Activity Type	To Learn	To Test
A) Normal Walk	3 R & 3 L.	3 R & 3 L.*
B) Running	2 R & 2 L.	2 R & 2L.
C) Backward walk	1 R & 1 L.	1 R & 1 L.
D) Walk with rotating arms	1 R & 1 L.	1 R & 1 L.
E) Side walk	1 R & 1 L.	1 R & 1 L.
F) Walk moving leg and arm on the same side together	1 R & 1 L.	1 R & 1 L.
G) Walk keeping touching wall	1 R & 1 L.	1 R & 1 L.
H) walk raising knees high	1 R & 1 L.	1 R & 1 L.
J) Gymnastic exercise	2	3

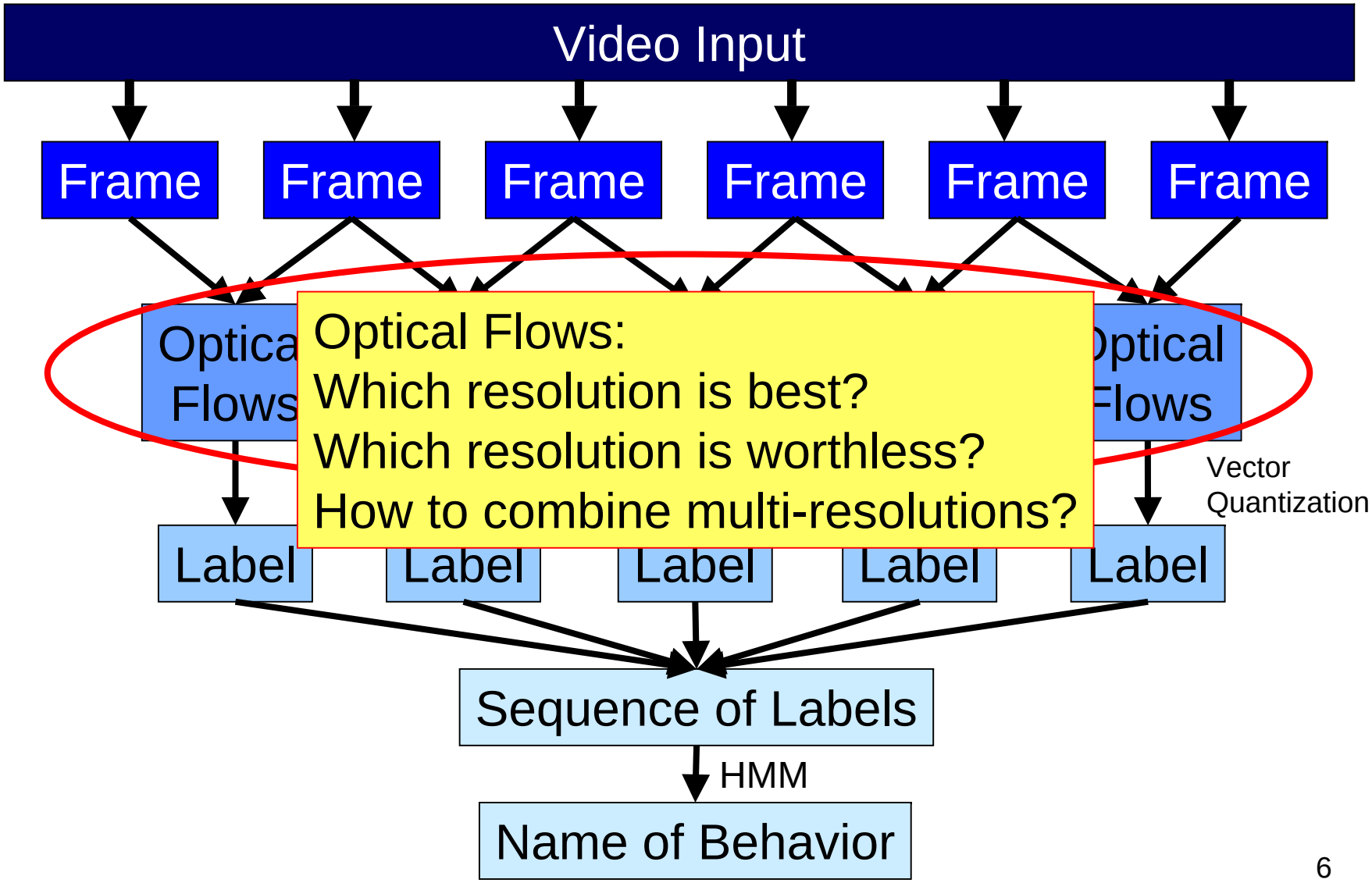
R: activity moving to right. L: activity moving to left.
* Four out of the 6 normal walk data are walks of a different person.

Example* of Recognition Result

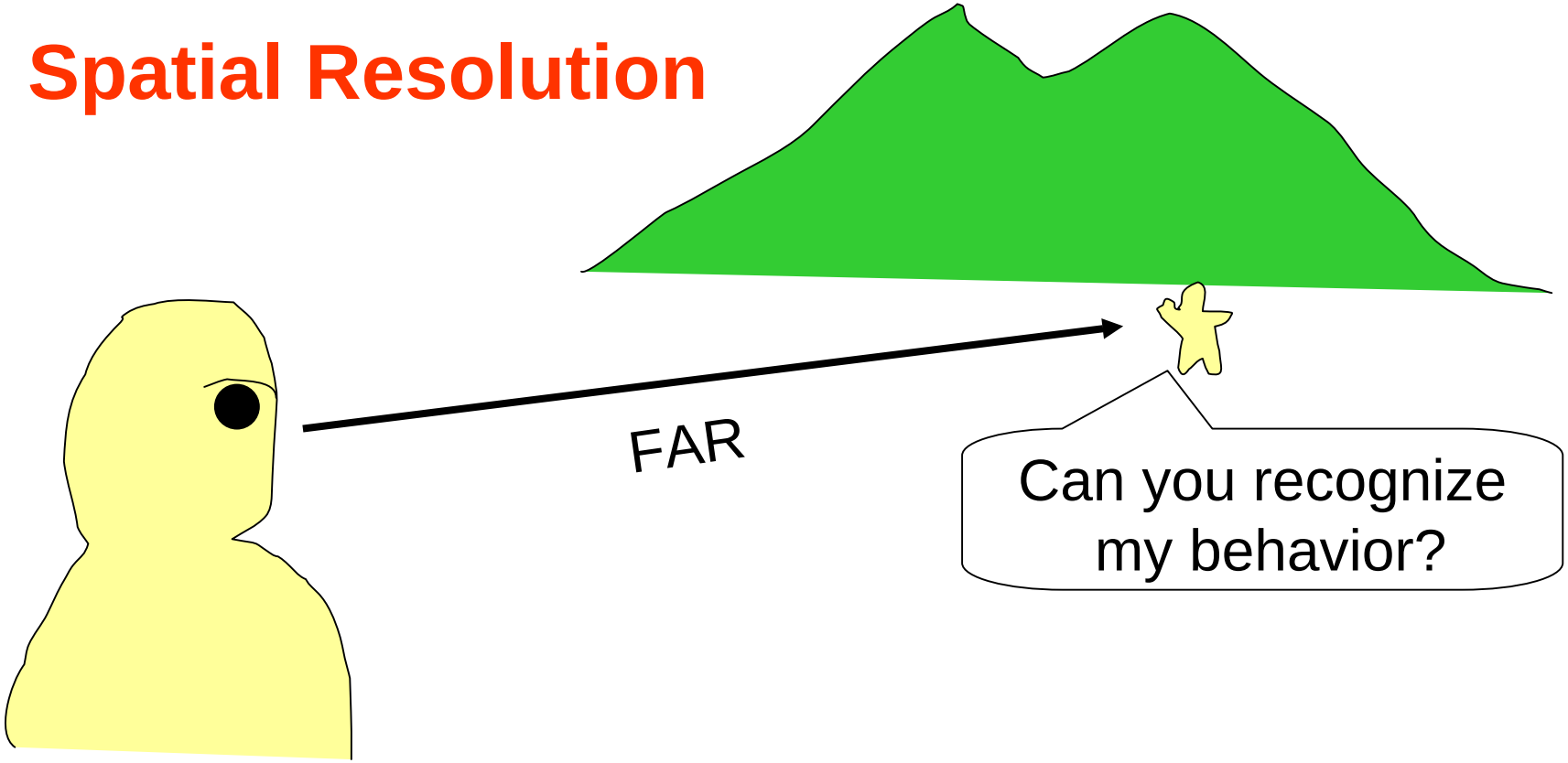
Behavior Type	1	2	3	4	5	6	7	8	9	Rate %
1) Normal Walk	4							2		66
2) Running		4								100
3) Backward Walk			2							100
4) Wk w/ rotating arms				2						100
5) Side Walk					2					100
6) Wk moving same side arm					1	1				50
7) Walk touching wall						1	1			50
8) Walk raising knee high								2		100
9) Gymnastic exercise									3	100

*Result may vary because of stochasity of HMM.

What is the Point?



Spatial Resolution

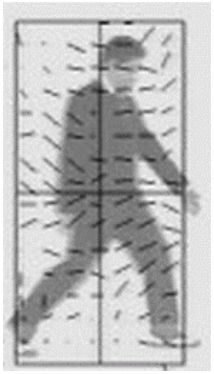


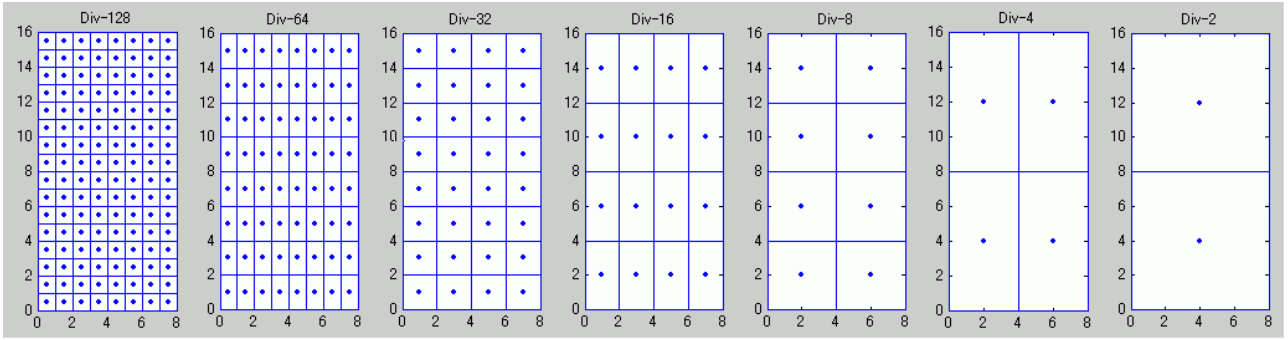
- Coarse resolution may be enough for action recognition.
- Most of information is optical flows.
- Not using 3D information.

Choice of Resolution Level

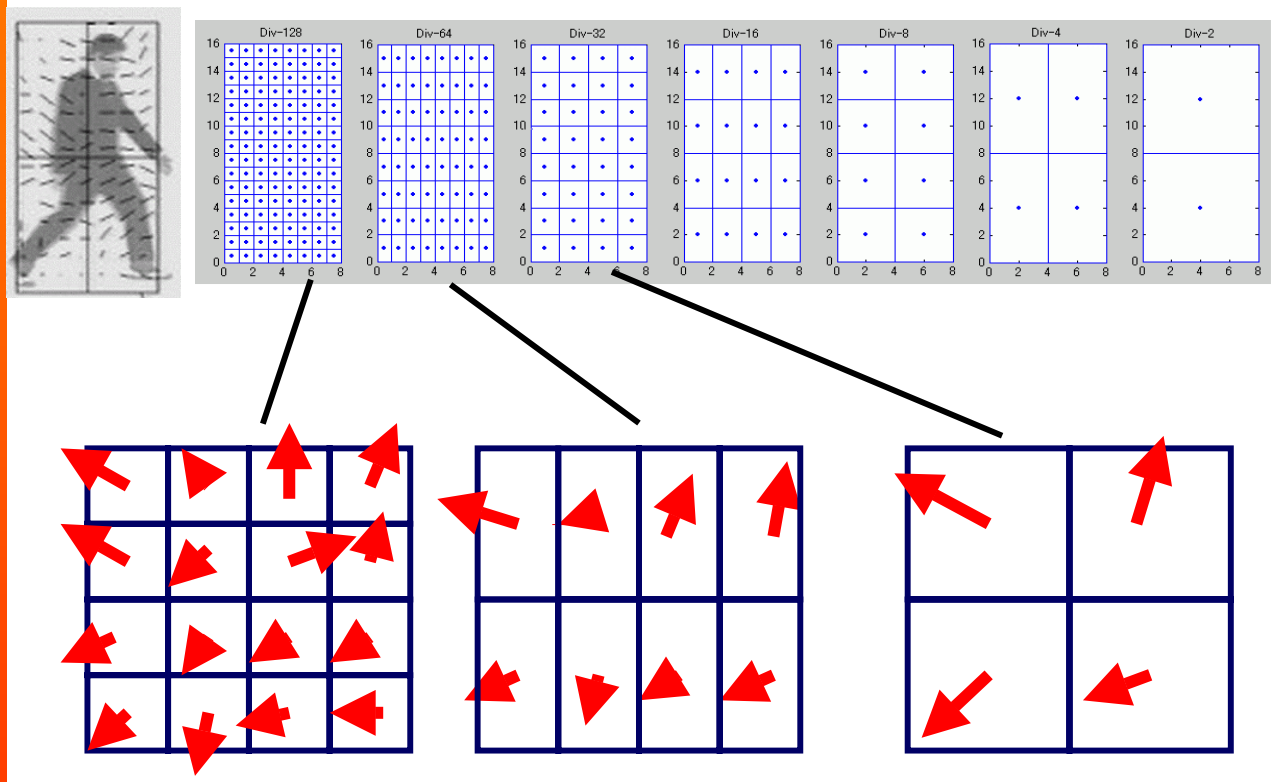
- Fine Reso: Large amount of information, but too local and Noisy.
- Coarse Reso: Small information, but relatively stable and global.

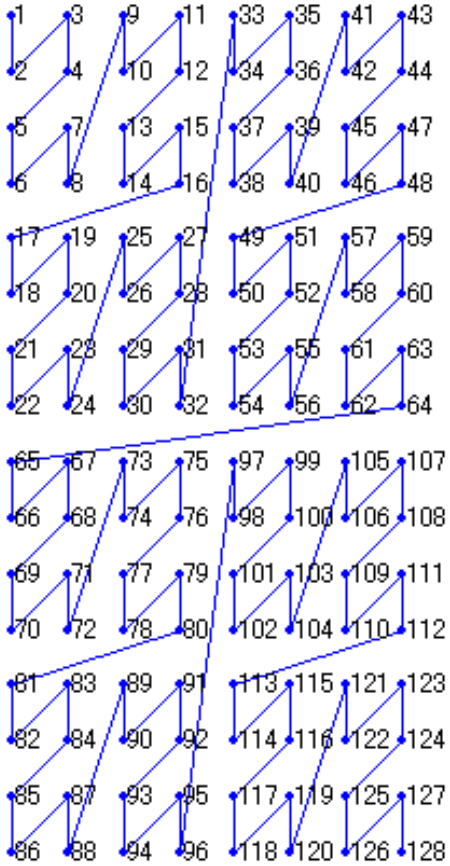
- Use only one resolution level?
- Or combine several levels?





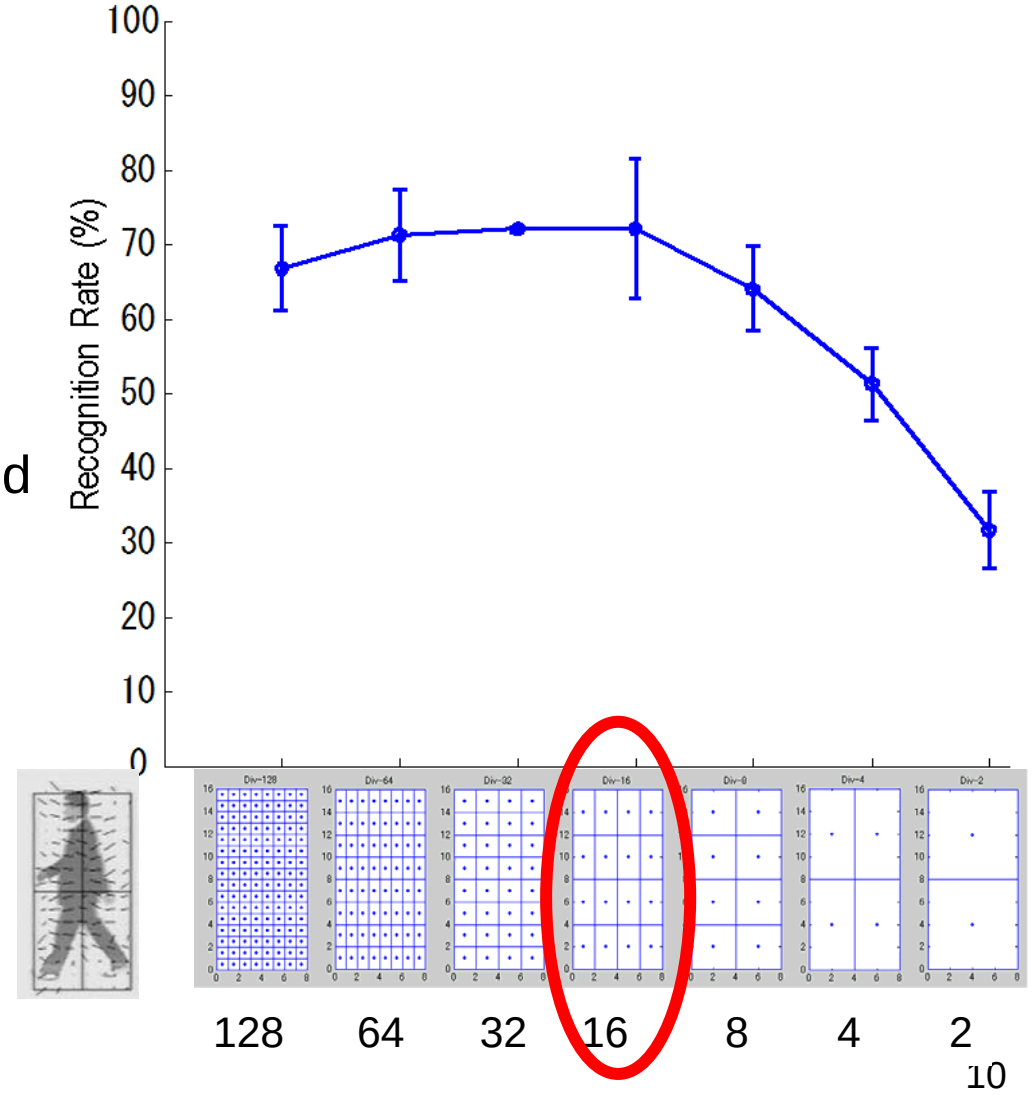
Fine to Coarse:
Local Sums of Optical Flows





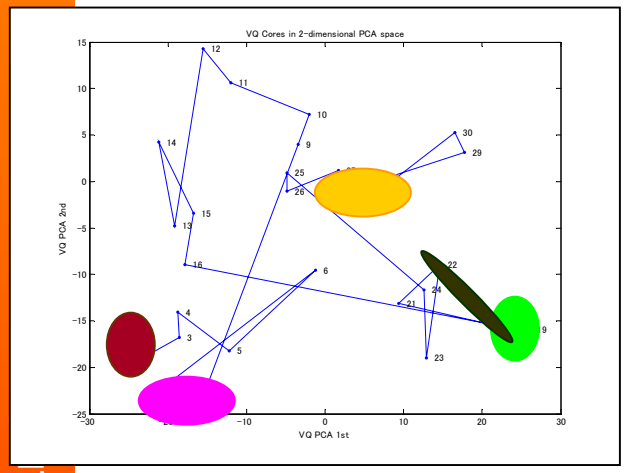
Recognition with one resolution level: Fine vs. Coarse

- Average recognition rates
  - over 10 times trials of HMM reconstruction.
- Good results at 16, 32, and 64 cells division.
  - 16 division is most efficient.

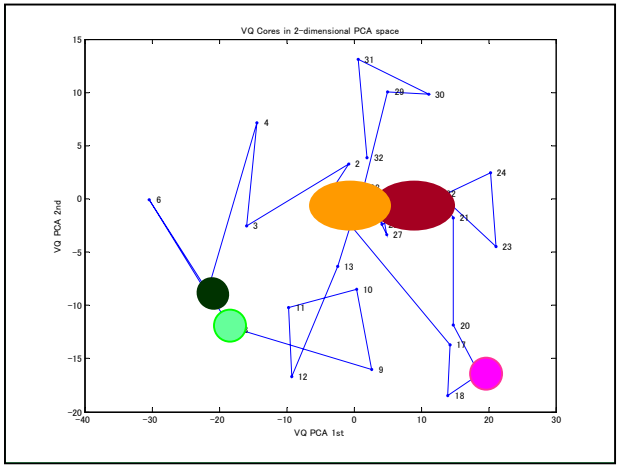


How did the system distinguished?

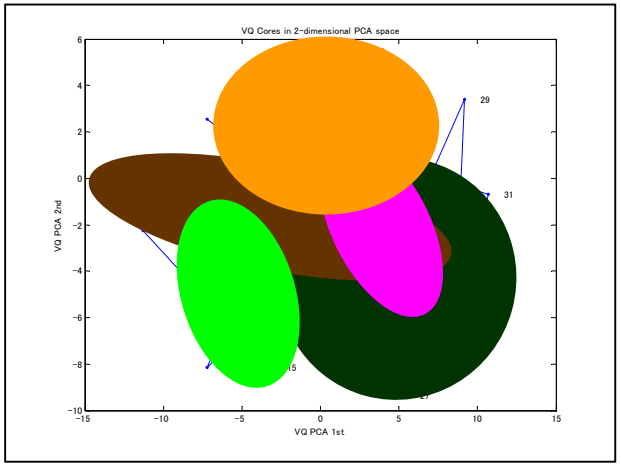
- Classification of Optical flows



At 128 division



At 16 division



At 2 division

Walk to left

Walk to right

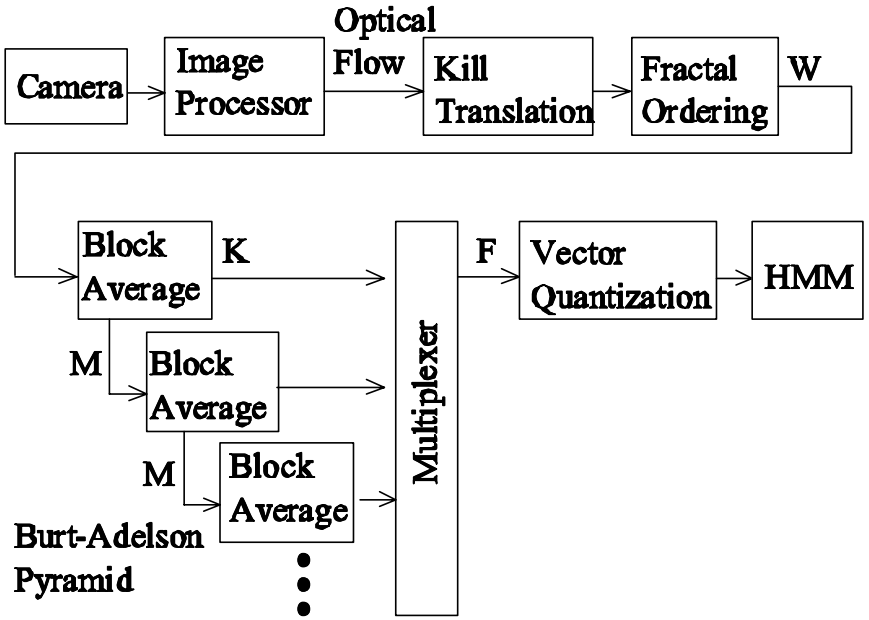
Gymn. Exer.

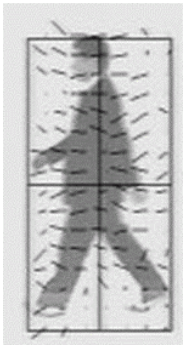
Run to left

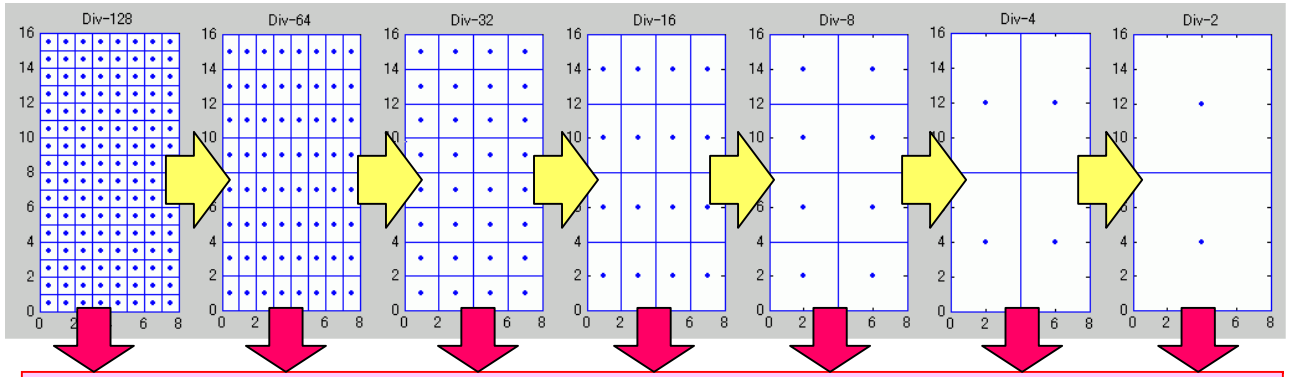
Run to right

Combination of multi-level resolution

- Burt-Adelson Pyramid

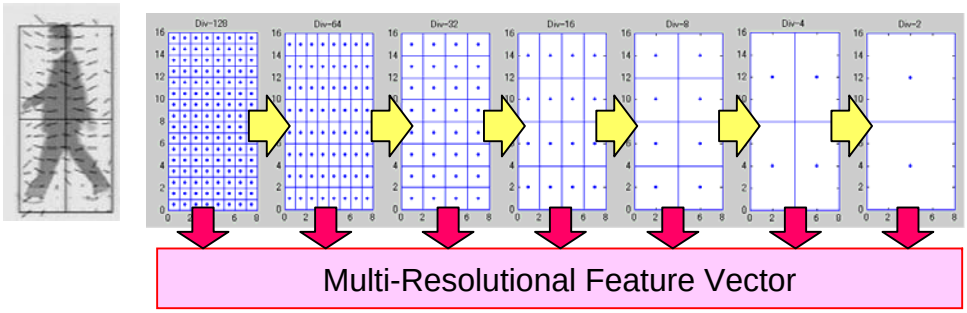


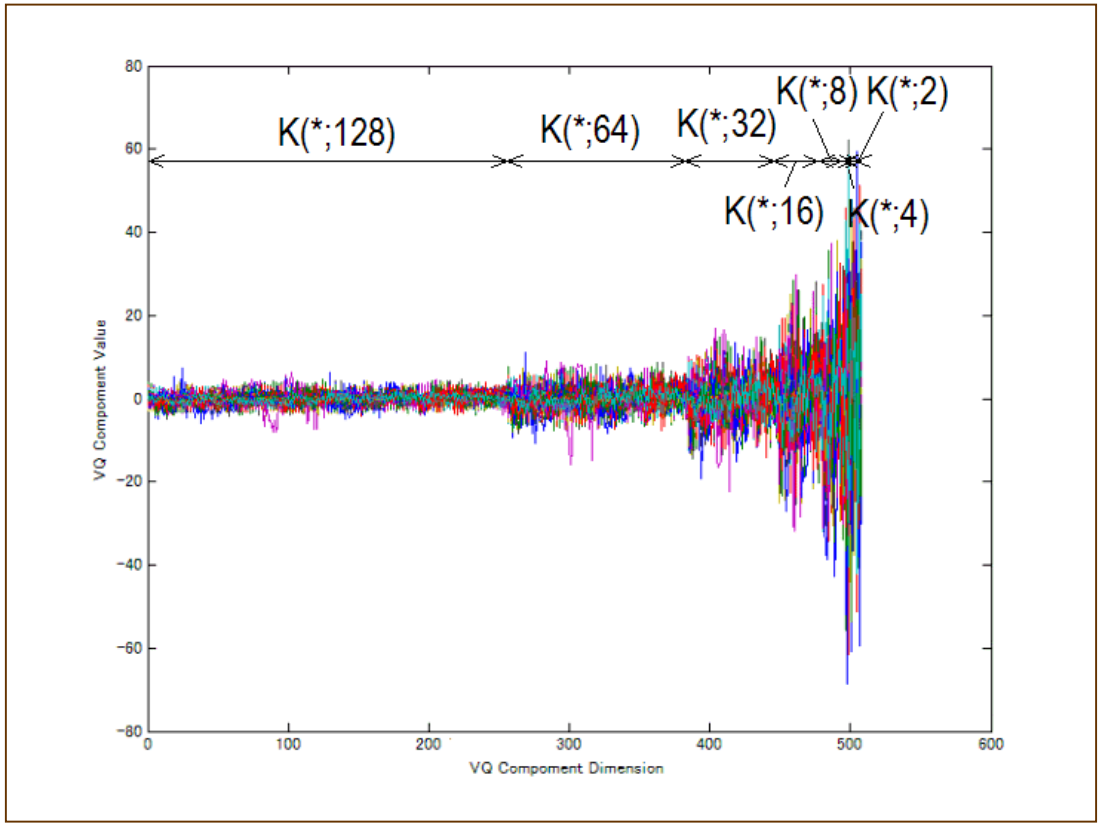




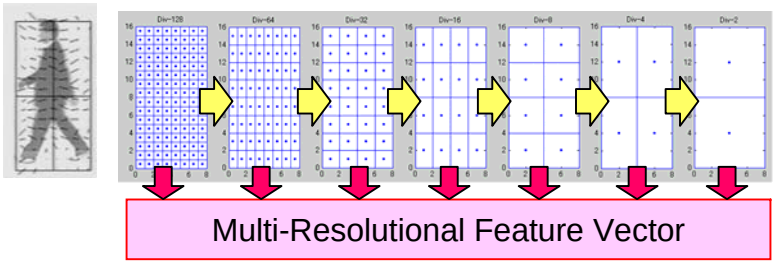
Multi-Resolutional Feature Vector

Constructing Multi-resolution Vector





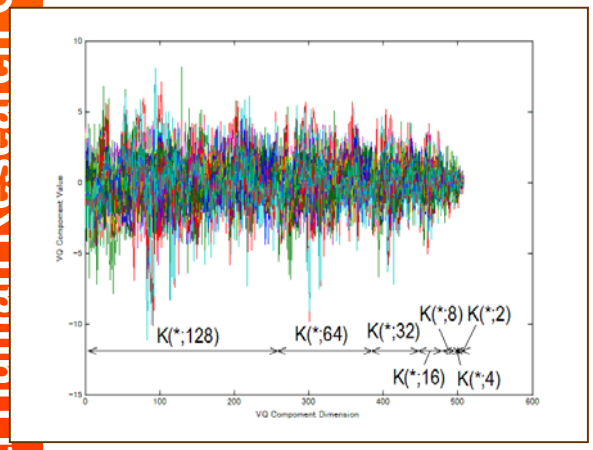
What weighting is best?



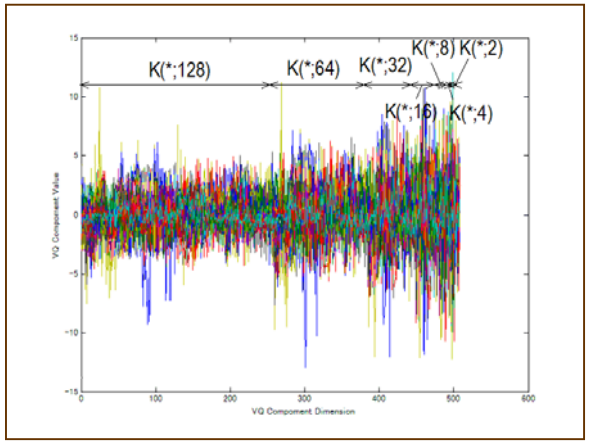
← Weight coefficients

$$A(D) = \begin{cases} 1/N & (\text{Block Ave.}) \\ \sqrt{D}/N & (\text{"Root" type}) \\ D/N & (\text{Block Sum}) \end{cases}$$

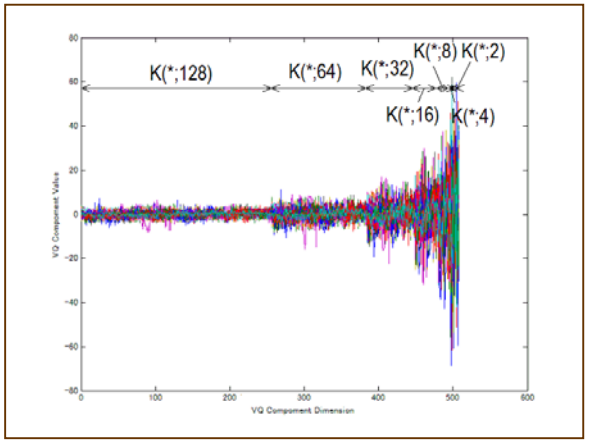
D: cell size, N: whole data size = 128.



Block Average

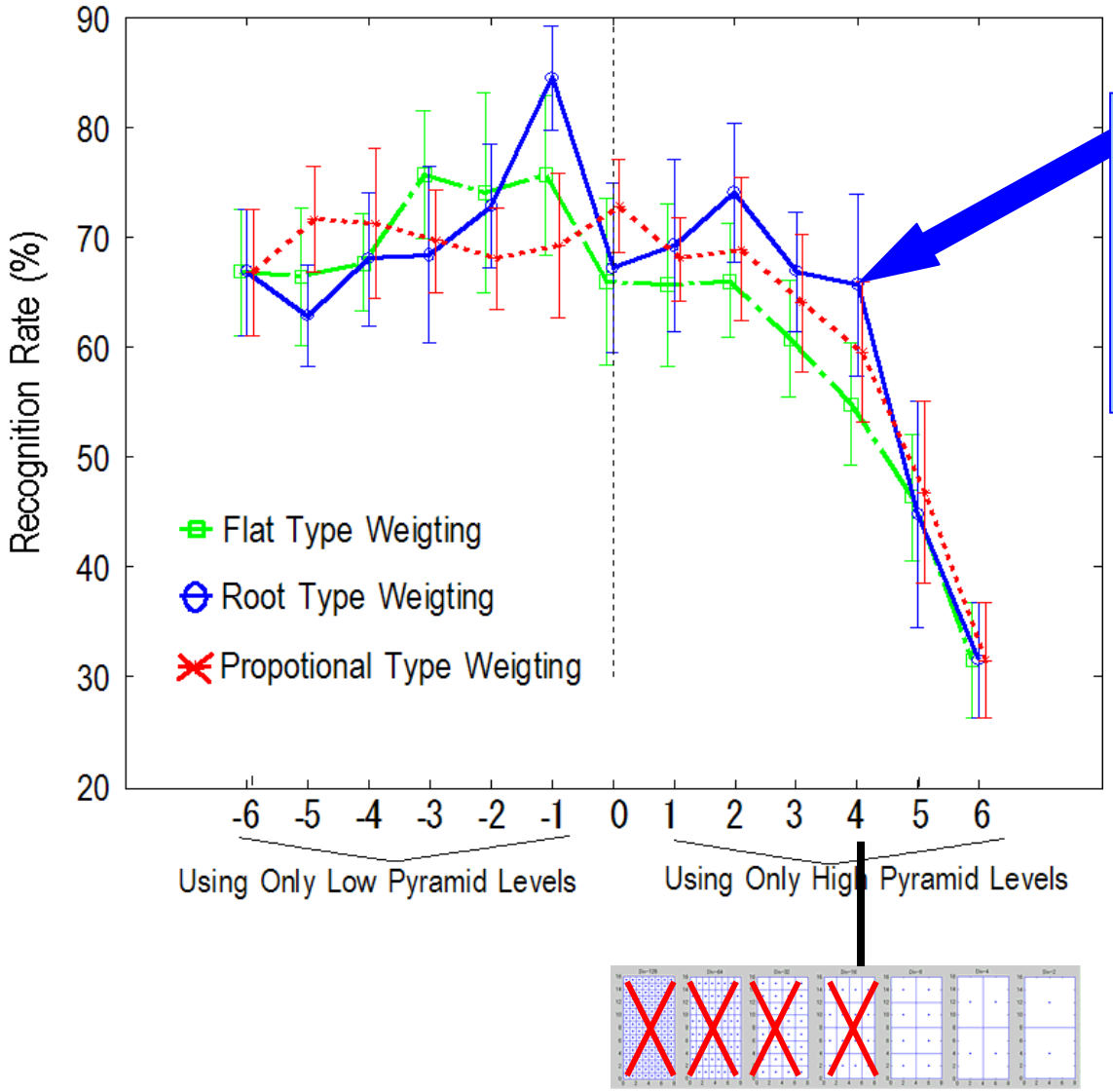


Root Type



Block Sum

"Root type" weighting is best!

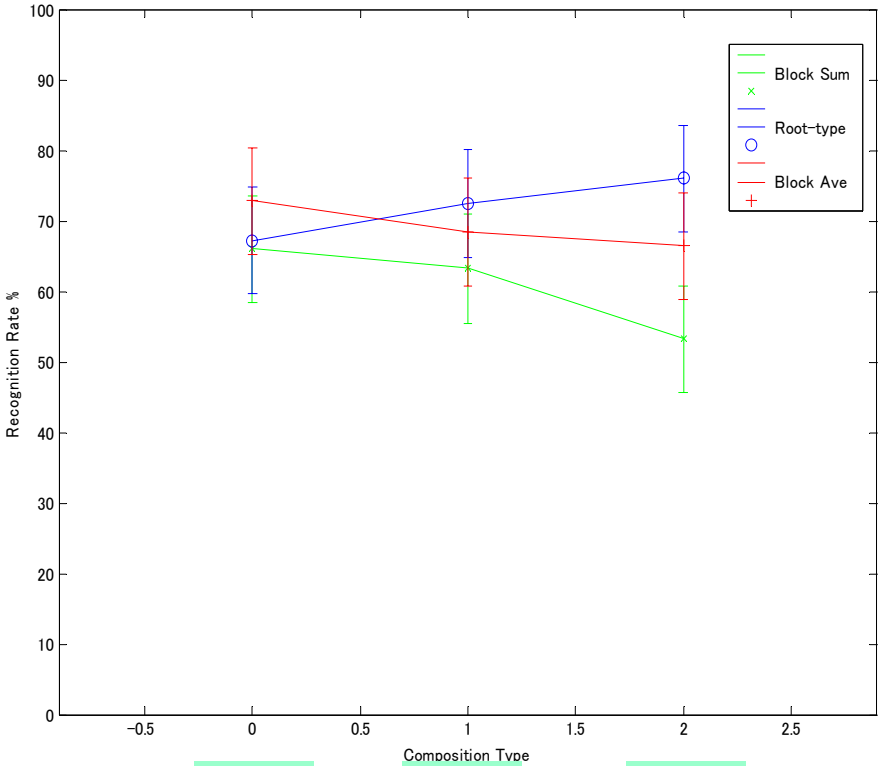


Root weighting using only 2-div, 4-div, and 8-div optical flows.

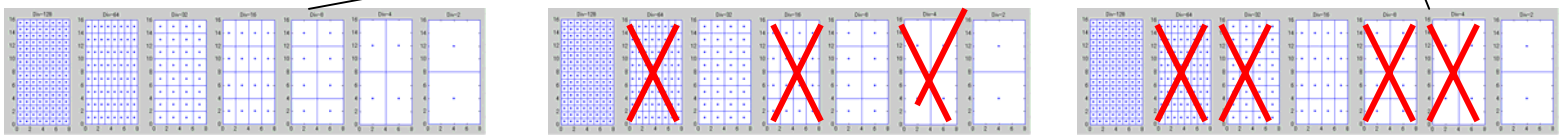
Sparse combination of levels

1. Combine all levels
   - 128, 64, 32, 16 ,8 ,4 ,2
2. Combine 4 levels
   - 128, 32, 8, 2
3. Combine 3 levels
   - 128, 16, 2

- Root type performs best.
- Block Sum is worst.



All 4/7 3/7



Conclusion

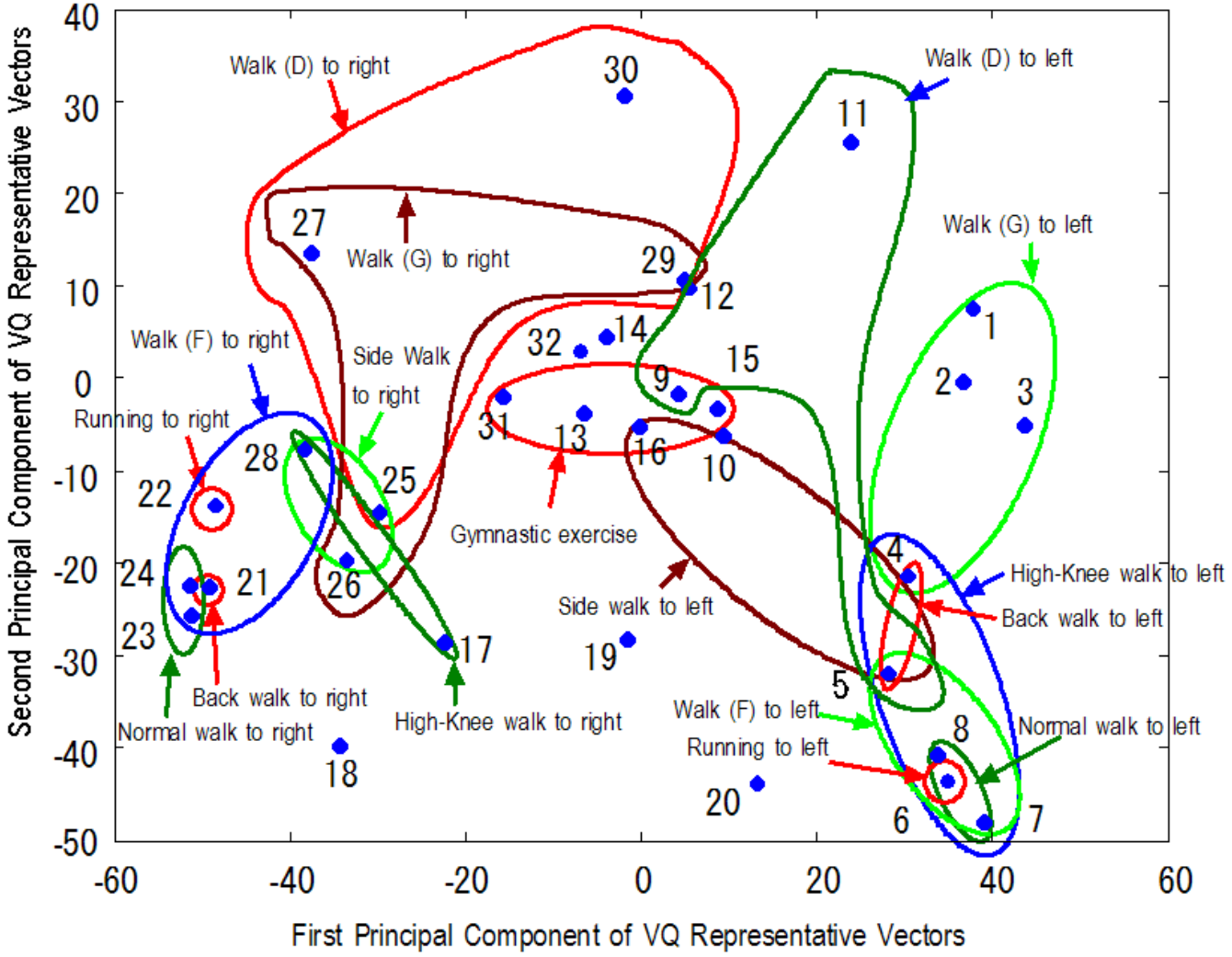
- Human Behavior recognition
  - Multi-resolution of optical flows
  - OF → VQ → HMM
- If using one level, which level is best?
  - 16 division.
- If using multi levels, which weighting is best?
  - Root weighting.

- 16 division: Such coarse visual information is enough for human activity recognition.

Digital Human Research Center, AIST

End

How the system distinguished



Root weighting. Using all levels of resolution

HMM Topology