Recognizing Human Activities in Video by Multi-resolutional Optical Flows

Toru Nakata

Digital Human Research Center, *National Institute of Advanced Industrial Science and Technology (AIST).*
Also Core Research for Evolutional Science and Technology (CREST), *Japan Science and Technology Agency (JST).*
Tokyo, Japan.

toru-nakata@aist.go.jp

Abstract **– A method to recognize human activities captured in video is proposed. The method classifies basic human body activities, such as walking, running, gymnastic exercises and others. Applying Burt-Adelson Pyramid approach, the system extracts useful features consisting of multi-resoutional optical flows. This paper also reports coarseness limit of spatial resolution of optical flow for activity recognition; optical flows of 8 sub-areas covering the human body area are minimum requirement for the recognition. Also, the experiment examines effective weighting of multi-resoutional feature components. These results on recognition of coarse video will be useful for designing surveillance camera system.**

Index Terms – Human activity recognition, Optical flow, Hidden Markov Model, Video Recognition.

I. Introduction

A. *Motivation*

Today, automatic and robust systems to recognize human activities for surveillance video camera are required strongly. Demand for security in public spaces and decline of camera price rushed us into employing huge number of surveillance cameras. However, number of personnel to monitor the cameras is insufficient. Camera systems should comprehend contents of the video images, especially human activities.

In general, motion pictures of surveillance video are coarse to recognize precise movements of human body, since areas of human body image are often small. Also, video is two-dimensional, so that it does not provide full information of actual three-dimensional movements. Therefore, it is essentially difficult to restore complete data of human body motion from surveillance video.

These difficulties should be overcome by developing methods that can recognize with information of small number of pixels. This paper proposes this kind of robust methods. I will explain my proposal focusing on the following 2 points:

1. Minimum requirement on number of pixels for human activity recognition.
2. Techniques of processing features to be efficient and robust for recognition.

B. *Biological background*

Animals and humans do not require so many pixels of motion image to recognize other animal motions. According to Johansson [1], visual information of movements of 12 major joints of a human body is enough for humans to recognize body activity.

Such coarse information is enough for recognition. Body movements are easy to comprehend, because they are strongly restricted. Body motions of animals are restricted by mechanical constrains. Also, animals and humans usually control their bodies with some typical coordination pattern [2]. Variety of biological movements is small.

This paper adopts optical flow of body images as cue for activity recognition instead of movements of the major joints as in the experiment of Johansson. Measurement of optical flows is easier than tracking movements of the major joints. Even though the optical flows do not represent detail of body motion exactly, they may provide enough amount of information for recognition.

C. *Related works*

There are 2 approaches in human activity recognition in video.

First approach is s*tructural movement analysis* represented by tracking interest points (that are often major joints) on a human body [3, 4]. Of course the movements of major joints well reflect posture and movement of a human body. However, it is rather difficult to track them stably when video is coarse or contains occlusions. Instead of ordinal interest points, Yilmaz and Shah [5] used singular points (and other geometrical features) of a spatiotemporal contour of human body.

The second approach is *statistical image processing* of frame images. Ma and Lin [6] processed frame images with primal component analysis (PCA) to classify and recognize sequence of human actions. Palana and Nelson [7] employed distribution of periodicities and magnitudes of body motions as features for recognition. These methods handle rather large amount of data, which are required for image processing.

Goa et al. [8] proposed an intermediate approach that analyzes optical flows statistically to comprehend structure of moving objects.

In general, conventional approaches require fine video images and complex calculation.

The approach of this paper can be classified into the statistical approach, because I do not track interest points on a human body. I concern realization of robust recognition methods that can work with coarse video images. According to the biological facts, statistical approaches may require fewer amount of information.

II. Method

A. Strategy Overview

My approach consists of 4 steps; 1) measuring optical flow over human area, 2) merging distribution of optical flows by using Burt-Adelson Pyramid method, 3) composing multi-resolutional feature vector that represent activity condition of each video frame, and 4) human activity recognition with Vector Quantization (VQ) and Hidden Markov Model (HMM).

Merger of neighboring optical flows is expected to have two advantages: reduction of data size and elimination of unimportant components. Well-planned spatial low-pass filters make image recognition successful. In general, image blurring makes information coarser, so it might affect badly for recognition. Blurring, however, eliminates unimportant information for recognition in some cases [9].

I employ Burt-Adelson Pyramid approach [10], which is a famous technique to analyze images by multi-scale features. The pyramid approach keeps information at several resolutional levels (called pyramid levels) to insure against danger of eliminating important components at the merging procedure.

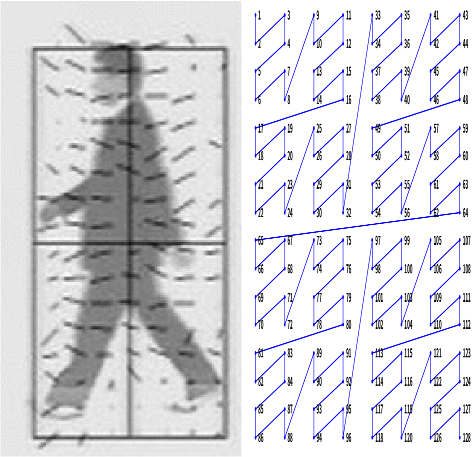


Fig. 1. Left: A sample of observed optical flows on a human body area. Right: Fractal numbering of observation points.

We do not know adequate cut-off frequency of the spatial low-pass filter, which changes as the case may be. So I employ this strategy.

B. Setting of Measurement

In order to make measurement stable, video recording condition in this paper is fixed as follows: number of human in video is assumed as one, and the distance between the human and the camera is almost constantly kept as 5 m (i.e. the human never comes closer or goes far). These assumptions are just for stable spotting of the human areas.

The system draws a rectangle over the human area (Fig.1). The center of the rectangle is set to the center of mass of pixels of the human body. The rectangle is posed to vertical and horizontal. Length of the vertical edges is fixed as 240 pixels, which is slightly larger than the human height. The length of the horizontal edges is fixed as 120 pixels.

The system then set a grid of observation points that has 128 points (in 8 columns and 16 rows) in the rectangle. For each observation points, the system calculates optical flow between frames by using template-matching method.

The system thus gets distribution of the 128 optical flow vectors denoted as V over the human area for each frame.

$$\vec{\mathrm{V}}_i(t) = (Vx_i(t), Vy_i(t)) \qquad (1)$$

For the aim of human activity recognition, we want to recognize human activity itself. The system should recognize walks (for example) by analyzing changes of actions inside human body area, even though the walks are slow or fast. Therefore, it may be better to ignore translational movements of the rectangle. The system uses optical flows without translational components described by (2).

$$\vec{\mathrm{W}}_i(t) = \vec{V}_i(t) - \frac{1}{N}\sum_{i=1}^{N}\vec{V}_i(t) \qquad (2)$$

where N is the number of the observation points that is 128.

C. Merger of Neighboring Optical Flows

The simplest merging of local data is block averaging, which just takes averages of data among neighbor observation points.

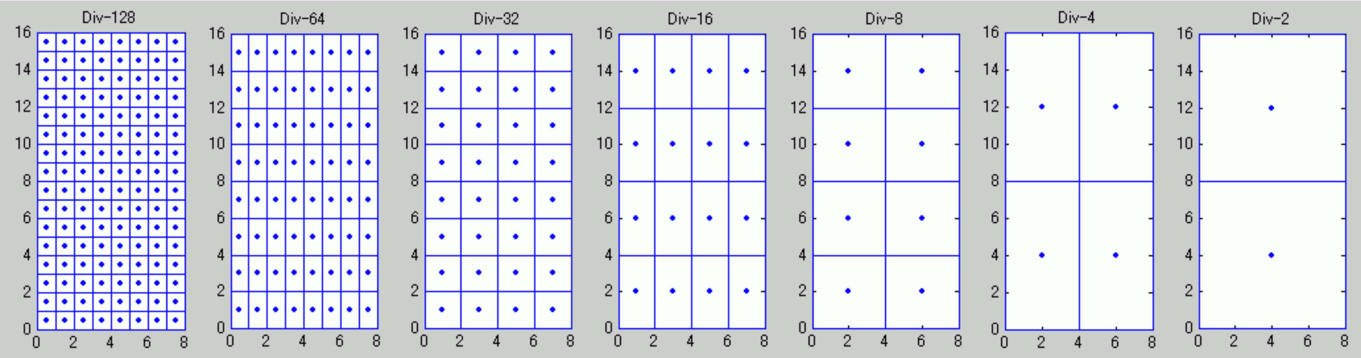


Fig. 2. Seven pyramid levels of optical flow merging.

Multi-scale block averaging proceeds by merging 2 neighboring cells recursively (Fig. 2). The fractal line shown in Fig.1 describes neighbor relationship of the merging. Recursive function of (3) and (4) generates the fractal numbering.

$$i = G(\log_2(N); x, y) \qquad (3)$$

$$G(\alpha; z, w) = \begin{cases} 1 & (\text{if } \alpha = 0) \\ G(\alpha-1; w, z) & (\text{if } w \leq 2^{\lfloor(\alpha-1)/2\rfloor}) \\ 2^{\alpha-1} + G(\alpha-1; w-2^{\lfloor(\alpha-1)/2\rfloor}, z) & (\text{otherwise}) \end{cases} \qquad (4)$$

where $\lfloor\ \rfloor$ stands for the floor function.

Using fractally numbered index instead of x-y coordinates, we can align the array of the optical flow vectors in a line that is convenient to calculate block averages.

$$\vec{M}(t; D, j) = \sum_{i=1+(j-1)N/D}^{jN/D} \vec{W}_i(t) \qquad (5)$$

M(t;D,j) of (5) is a block summation of optical flows in the j-th sub-area, when the rectangle area is divided into D parts.

K of (6) is a composition of M vectors and holds the input information that is low-pass-filtered at a certain resolution level.

$$K(t; D) = \left\{\vec{M}(t; D, 1), \cdots, \vec{M}(t; D, D)\right\} \qquad (6)$$

We can select the number of the sub-areas, D. D should be a power of 2.

$$D = 2^{7\text{-L}} \qquad (7)$$

We call the number of L pyramid level.

D. *Composing Scale-Free Feature Vector*

To maintain enough information for recognition, it is better using several K vectors at different pyramid levels than using single K. The system composes a feature vector, F, that consists of K vectors (7).

$$F(t) = \bigoplus_{L=0}^{6} \{A(D)K(t; D)\} \qquad (7)$$

where A is a weight term for the multiplication. The dimension of F becomes 510 when N is 128.

There are three practical alternatives for style of the weight term.

$$A(D) = \begin{cases} 1/N & (\text{"Flat" type}) \\ \sqrt{D}/N & (\text{"Root" type}) \\ D/N & (\text{"Propotional" type}) \end{cases} \qquad (8)$$

Flat type of the weight attaches big decisiveness for the recognition on global components (that are components of higher pyramid levels). When there exist some correlations among the W vectors (the amended optical flows), global components grow large (as shown in Fig. 6).

Root type is expected equalizing norms of K vectors, when optical flows are random (so that the merged optical flows M distribute like random walk and will obey (9)).

$$\left|M(t; D_{small}, 1)\right| \propto \left|M(t; D_{\text{large}}, 1)\right| \sqrt{\frac{D_{\text{large}}}{D_{small}}} \qquad (9)$$

Proportional type makes the feature vector an alignment of the block averages ignoring sizes of the blocks. It will give detail components more decisiveness.

E. *Recognition Procedure*

We can use ordinal techniques to recognize human activities that are presented as time-series of the feature vector.

First, we employ Vector Quantization technique to classify F vectors. The system calculates F vectors of all time frames of all videos reserved for learning (as shown in Table I) and classifies them into 32 groups.

At this step, a video clip is processed into a sequence of numbers that stand for types of distribution of the optical flows.

We then employ Hidden Markov Model to recognize each type of human activity. At learning step, the system generates a state-transition model for each human activity. State transition topology of the model is shown in Fig. 3. It is common to all kinds of the human activity samples.

After learning, the system recognizes human activities by using acquired models.

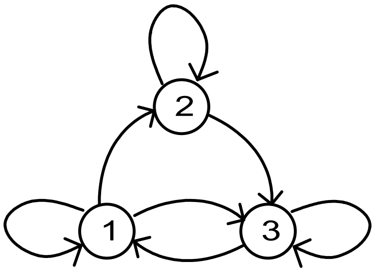
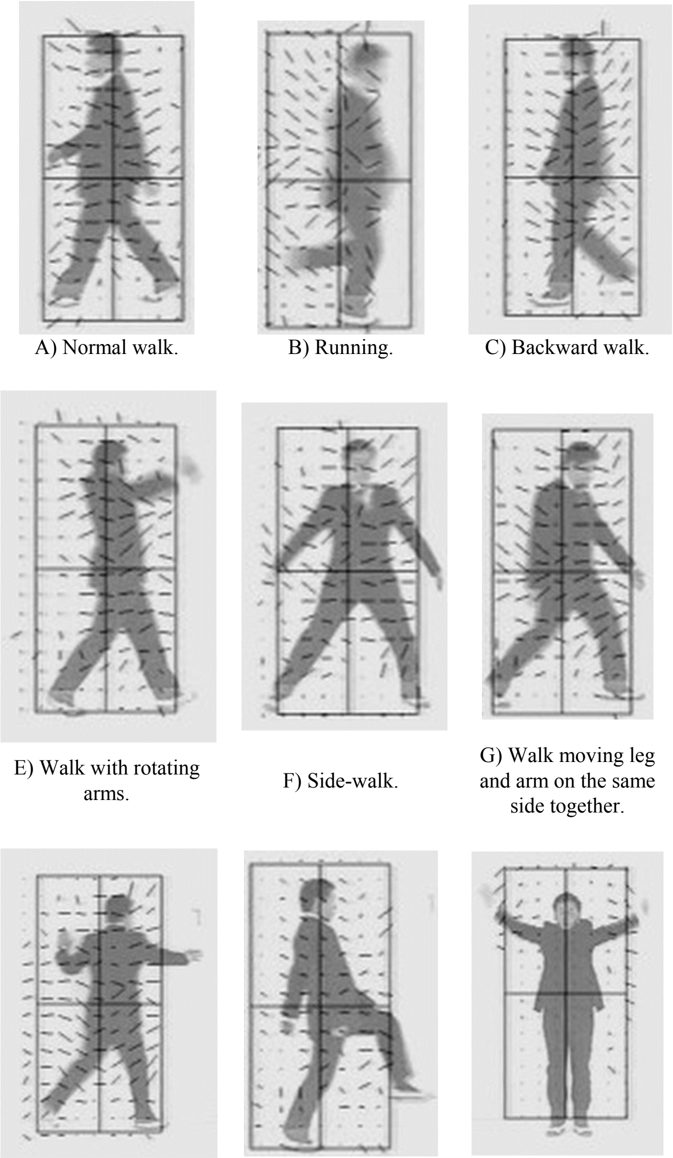


Fig. 3. Topology of the transition matrix used in Hidden Markov Model Process.

TABLE I
NUMBER OF SAMPLE DATA FOR LEARNING AND RECOGNITION TEST

Activity Type	For Learning	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.
E) Walk with rotating arms	1 R & 1 L.	1 R & 1 L.
F) Side walk	1 R & 1 L.	1 R & 1 L.
G) Walk moving leg and arm on the same side together	1 R & 1 L.	1 R & 1 L.
H) Walk keeping touching wall	1 R & 1 L.	1 R & 1 L.
I) walk raising knees high	1 R & 1 L.	1 R & 1 L.
J) Gymnastic exercise	2	3

R: activity going to right. L: activity going to left.



A) Normal walk.

B) Running.

C) Backward walk.

E) Walk with rotating arms.

F) Side-walk.

G) Walk moving leg and arm on the same side together.

H) Walk keeping touching wall.

I) Walk raising knees high.

J) Gymnastic exercise.

Fig. 4. Activities and optical flows for the experiment.

III. Experiment

A. *Observation Setup and Input Data*

I used a Victor GR-HD1 camera that provides monochrome NTSC signals of video image. The image processing equipment is a Hitachi HICOS IP7000, which takes 77 ms on average to calculate 128 optical flows of each frame. Therefore the system gets about 13 image frames for each second.

Videos of human activities to learn and recognize consist of 9 kinds of behaviors shown in Fig. 4. Half of the video samples are used to learning of VQ and HMM. The rest half of the video samples are for recognition test (Table I). All

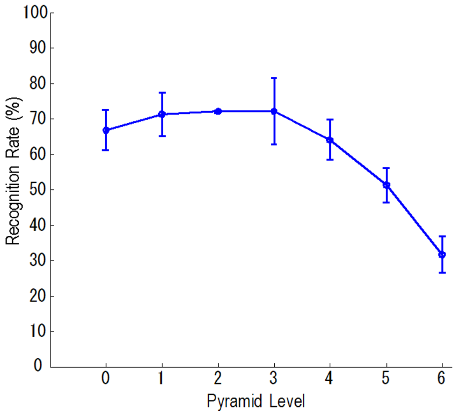


Fig. 5. Pyramid levels versus average recognition rate. Error bars stand for standard deviations.

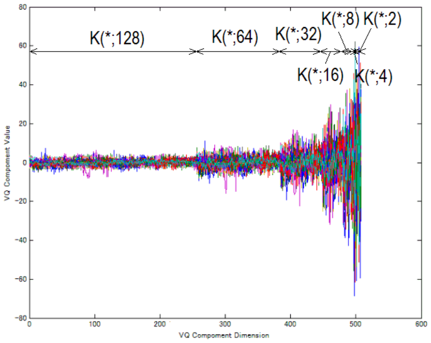


Fig. 6. Spectra of the 32 code vectors representing the feature vectors.
(Weight type of feature vector is flat type.)

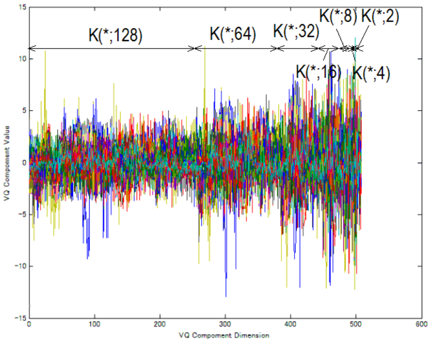


Fig. 7. Spectra of the 32 Code vectors representing the feature vectors.
(Weight type of feature vector is root type.)

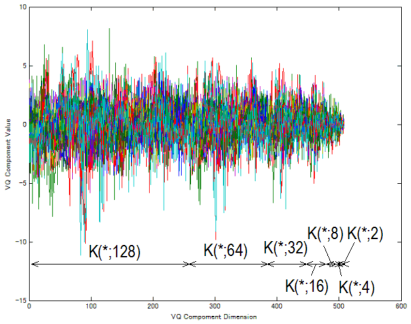


Fig. 8. Spectra of the 32 Code vectors representing the feature vectors.

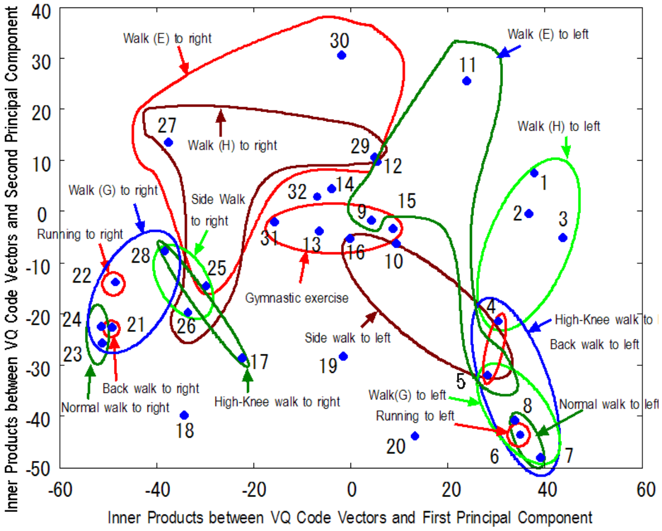


Fig. 9. Distribution of representative feature vectors classified by vector quantization. (Number of code vectors = 32, Weight type of feature vector = root type. The feature vectors are positioned by primal component analysis.)

videos record behaviors of same person, except 4 samples in walk videos.

The person of the videos wears black clothes and moves in front of a white wall. So the system can extract pixels on the person body by subtracting initial image of the white wall.

I conduct the HMM procedure 10 times for each pyramid level, since performance of the HMM recognition is unstable due to dependence of randomization of initial variables.

B. *Result and Discussion*

1) Recognition with single pyramid level:

The system recognized the human activities using just single pyramid level at the success rates shown in Fig. 5.

Till the pyramid level become 3, at which the whole image is divided into 16 sub-areas, the system could recognize at the rates around 70%. The feature vector when the pyramid level is 3 has only 16 dimensions and got similar result to the result of pyramid level of 0 whose the feature vector has 128 dimensions.

2) Recognition with multi pyramid level:

I also conducted an experiment to recognize human activities with feature vectors including several pyramid levels.

The vector quantization procedure classified the feature vectors into 32 representatives whose spectra are shown in Fig. 6, 7, and 8 where the weightings of feature components are flat type, root type, and proportional type respectively. As expected, the root-type weighting realized equalization of norms of the feature vector.

Each of the representative vectors correspond to one of typical frames of the human activities (Fig.9). The vector quantization processed the video sequences into sequences of indexes of the representative vectors (Fig.10, Fig.11).

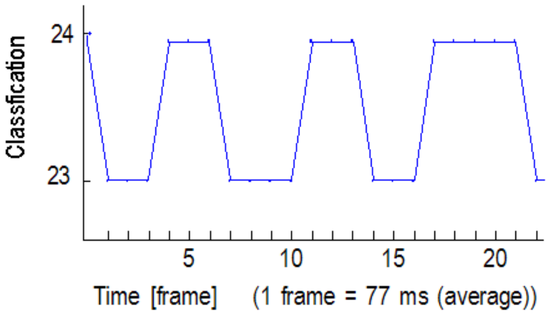


Fig. 10. Temporal transition of frame classification during one of the walk samples. Each transition stands for a walk step.

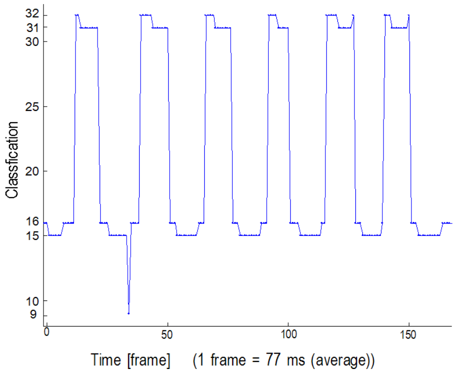


Fig. 11. Temporal transition of frame classification during one of the gymnastic exercise samples. Each Transition stands for alternation of the body posture.

The recognition rates were not so different when the system used all of 510 components of the feature vectors (Fig. 12).

Since 510 dimension of the feature vector is large, I tried running the system with dimension-reduced feature vectors. I discarded components of some pyramid levels (Table II).

As the result, I found that the root type weighting got better and stable recognition rate when the feature vector has only components of high pyramid levels up to 4. It means that the system can recognize the human activities with coarse information of 8 sub areas as well as recognition using detail video information.

3) Recognition with sparse pyramid level:

I conducted recognition tests to find adequate composition of pyramid levels to perform more efficient recognition, which get high success rate with fewer amount of information.

I selected pyramid levels of the feature vector to reduce the data size (Table III). Despite the reduction of information, the system got almost same success rates as that of normal setup (Fig.13).

Root type weighting achieved relatively better success rate again.

TABLE II
COEFFICIENTS FOR PARTIAL USE OF THE FEATURE VECTOR

X of Fig.	Pyramid level						
	0	1	2	3	4	5	6
-6	A(128)	0	0	0	0	0	0
-5	A(128)	A(64)	0	0	0	0	0
-4	A(128)	A(64)	A(32)	0	0	0	0
-3	A(128)	A(64)	A(32)	A(16)	0	0	0
-2	A(128)	A(64)	A(32)	A(16)	A(8)	0	0
-1	A(128)	A(64)	A(32)	A(16)	A(8)	A(4)	0
0	A(128)	A(64)	A(32)	A(16)	A(8)	A(4)	A(2)
1	0	A(64)	A(32)	A(16)	A(8)	A(4)	A(2)
2	0	0	A(32)	A(16)	A(8)	A(4)	A(2)
3	0	0	0	A(16)	A(8)	A(4)	A(2)
4	0	0	0	0	A(8)	A(4)	A(2)
5	0	0	0	0	0	A(4)	A(2)
6	0	0	0	0	0	0	A(2)

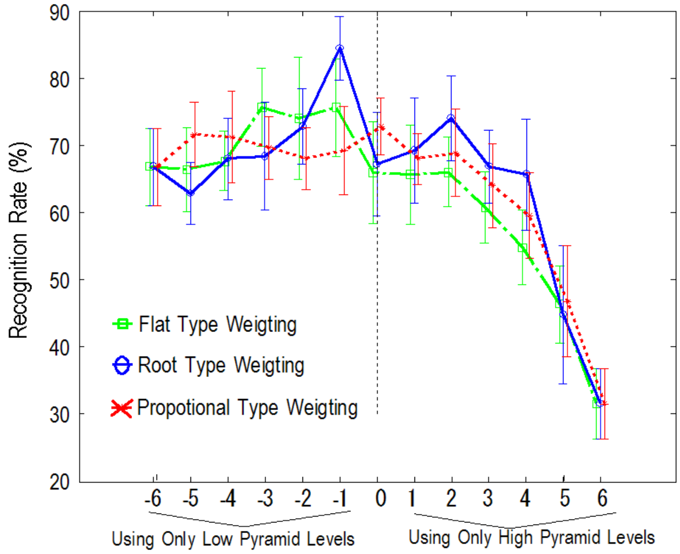


Fig. 12. Composition of different pyramid level features versus recognition rate.

IV. CONCLUSION

This paper proposed a method to recognize human activities in video by processing optical flows with approaches of Burt-Adelson pyramid, vector quantization and Hidden Markov Model.

There are 2 findings from the trial:

1) optical information of 8 sub-areas shooting over a human whole body is enough to saturate recognition rate of this method.
2) root type weighting is most effective to make the recognition robust.

In future works, the method should be utilized to recognize other human behavior, especially movements of approaching or retreating against the camera and small behaviors within a part of a human body.

REFERENCES

TABLE III
SPARSE FEATURES AND RECOGNITION RATE

Type	Coefficients for pyramid levels							Rate	Stdev
	0	1	2	3	4	5	6	(%)	(%)
All-flat	1	1	1	1	1	1	1	66.0	7.6
All-root	1	a	a^2	a^3	a^4	a^5	a^6	67.2	7.7
All-prop	1	a^2	a^4	a^6	a^8	a^10	a^12	72.8	4.1
4/7-flat	1	0	1	0	1	0	1	63.2	7.7
4/7-root	1	0	a^2	0	a^4	0	a^6	72.4	5.5
4/7-prop	1	0	a^4	0	a^8	0	a^12	68.4	4.0
3/7-flat	1	0	0	1	0	0	1	53.2	7.6
3/7-root	1	0	0	a^3	0	0	a^6	76.0	5.0
3/7-prop	1	0	0	a^6	0	0	a^12	66.4	3.9

* a = sqrt(1/2).

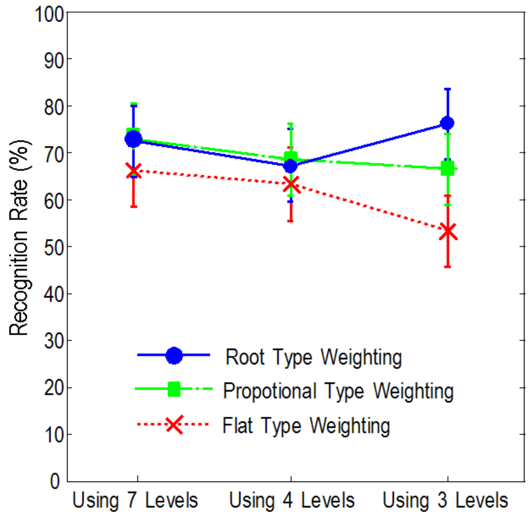


Fig. 13. Composition of sparse features versus recognition rate.

[1] G. Johansson, Visual motion perception, *Scientific American,* vol.232, pp.76-88, 1975.

[2] A. A. Sharp, E. Ma, and A. Bekoff, Developmental changes in leg coordination of the chick at embryonic days 9, 11, and 13: uncoupling of ankle movements, *J. Neurophysiology*, vol. 82 no. 5, pp. 2406-2414, 1999.

[3] C. M. Lu and N. J. Ferrier, Repetitive motion analysis: segmentation and event classification, *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 2, pp. 258-263, 2004.

[4] A. Yilmaz and M. Shah, Recognizing human actions in videos acquired by uncalibrated moving cameras, IEEE ICCV, 2005.

[5] A. Yilmaz and M. Shah Actions Sketch: A Novel Action Representation, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2005.

[6] G. Ma and X. Lin, Typical sequences extraction and recognition, *HCI/EECV2004*, pp.60-71, Springer-Verlag, 2004.

[7] R. Polana and R. C. Nelson, Detection and recognition of periodic, non-rigid motion, *Intl. J. Computer Vision,* vol. 23, no. 3, pp. 261-282, 1997.

[8] J. Gao, R. T. Collins, A. G. Hauptmann and H. D. Wactlar, Articulated motion modeling for activity analysis, CIVR 04, 2004.

[9] H. Fujisawa and C. L. Liu, Directional pattern matching for character recognition revisited, IEEE ICDAR 2003, 2003.

[10]P. J. Burt and E. H. Adelson, The Laplacian pyramid as a compact image code, *IEEE Trans. Communication,* vol. 31, no. 4, pp. 532-540, 1983.