Sparse Representation of Adaptive Key Frame Features for Human Action Classification

Kanokphan Lertniphonphan[1], Supavadee Aramvith[2], Thanarat H. Chalidabhongse[3]
[1,2]Department of Electrical Engineering, Chulalongkorn University, Bangkok, Thailand
[1]E-mail: Kanokphan.l@student.chula.ac.th
[2]E-mail: Supavadee.A@chula.ac.th
[3]Department of Computer Engineering, Chulalongkorn University, Bangkok, Thailand
[3]E-mail: Thanarat.C@chula.ac.th

***Abstract*— Human action movement has constrained by the articulated body which leads to the variation of movement velocity from point-to-point. In this paper, adaptive key frame intervals are used to specify the proper number of frames by detecting the variation of human motion. Features which are extracted within the interval contain information of primitive movement which is similar among the same action. Then, the sparse representations of primitive movement are trained. The results on WEIZMANN demonstrate that the sparse representation within adaptive key frame interval can effectively classifies actions.**

I. INTRODUCTION

In the recent years, many applications, such as automated surveillance, human computer interaction interface, and video indexing, consider visual based human action classification as one part of the system to improve the efficiency and to reduce the workload of human. Several feature representations of human action classification are reviewed in [1] and [2]. The challenges of human action classification are from several factors such as appearance, posture, moving speed, and moving direction. The selected feature representation is considered based on the above factors.

Features for human action can be categorized into appearance-based and motion-based. The appearance-based feature considers human posture by extracting the information from silhouette. In [3], the self-similarity of the foreground silhouette was used to detect periodic motion. Histogram of oriented gradient (HOG) was used to recognize the primitive pose in still images in [4]. In [5], contour of silhouette was used as a pose representation. For the motion-based features, the features focus on the information in temporal domain. Several methods [6], [7], constructed motion descriptor based on optical flow computation in the consecutive frames. Bradski and Davis [8] introduced time Motion History Image (tMHI) which contains layered silhouettes in a period of time to extract motion gradient. In addition, there are several approaches that extract information in spatio-temporal domain. The work in [9] introduced cuboids which were detected by using the separable linear filter in the spatio-temporal. In [10], the slow feature was extracted by using slow feature analysis on random sampling cuboids. There are several approaches that extracted features in every frame [11], in periodical interval [12], [13], or in the entire sequence [14], [7]. However, a problem of constructing compact and useful features is still remained.

For extracting compact and informative action representation, there are several researches which classified actions based on key frame features [15], [16], [17], [18], which are the discriminative features for human action classification. In [19], human actions were recognized by a small number of frames. The extracted key frames represented the characteristic among actions. The key frame was used to specify a number of frames for extracting features [20]. The sparse representation was used in a coding scheme to find the accurate and compact action representation in [21]. The sparse representation is widely used in computer vision applications since its representation and learning can be used to reconstruct the large dimension data by using a few basis [22].

In this work, we considered in feature representation which contains information in spatio-temporal domain. By finding key frames, the extracted features within a key frame interval, are used to construct history images. The history images consist of Adaptive Motion History Image (AMHI) and Key Pose Energy Image (KPEI) [20]. Also, the period of extracted window is not fixed but varies based on the number of frames in key frame interval which can be automatically adapted. Then, sparse representation are constructed from the histogram of the history images to classify action on WEIZMANN [12] dataset.

This paper is organized as follows. The overview of our system is explained in Section II. Feature extraction and action classification are presented in Section III and Section IV, respectively. Finally, the experimental results are discussed in Section V.

II. SYSTEM OVERVIEW

The processes of action classification are shown in Fig. 1. The system starts with human segmentation from background. Then, a key frame is detected based on motion occurrence in the consecutive frames. At each key frame, the history images, which are AMHI and KPEI, are constructed by layered silhouettes in key frame interval. The images contain both motion direction and posture transformation information of an action cycle.

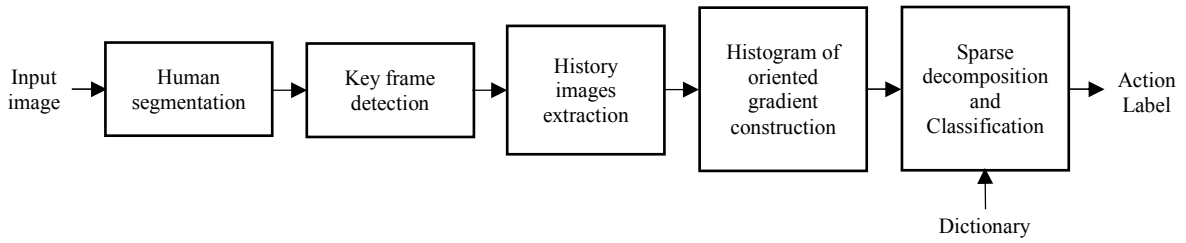

Fig. 1. System overview.

The oriented gradient of the history images are computed and used to construct histogram as an action descriptor for each key frame interval. The sparse representation of the feature descriptor is used to classify action by finding the minimum residual from the dictionary.

III. FEATURE EXTRACTION

For feature extraction, we start with key frame extraction since WIEZMANN dataset [12] provides foreground mask for segmentation. In this section, finding adaptive key frame interval, AMHI and KPEI extraction, and image oriented gradient histogram construction are described.

A. Adaptive Key Frame Interval

The concept of history images AMHI and KPEI are similar to tMHI [8]. tMHI represents motion gradient from successive layer of silhouette. Pixel intensity is based on the recentness of the layer. Given the current time stamp t, and the maximum time of updating period δ, the updating function of $tMHI_\delta(x,y)$ is defined as in (1).

$$tMHI_\delta(x,y) = \begin{cases} t\ if\ current\ silhouette\ at\ (x,y) \\ 0\ else\ if\ tMHI_\delta(x,y) < (t-\delta) \end{cases} \quad (1)$$

The limitation of a fixed updating period is that the extracted features are not invariant to speed variation which bases on performer and action. So the adaptive interval for extracting history images, are used to extract a cycle of primitive movement.

In each input image, key frame is detected by using foreground and motion information [20]. The foreground can be extracted by background subtraction. The motion region is computed by using frame differencing in the consecutive frames. Key frame is detected by computing an amount of motion variation over the action cycle. At time t, the preprocessed binary foreground image $F_t(x,y)$ and the binary motion frame differencing image $D_t(x,y)$ are used to find variation in motion. The pixels in the foreground region and the moving area are counted. The change in motion at each frame M_t is defined as in (2).

$$M_t = \frac{\sum_{(x,y)\in I} D_t(x,y)}{\sum_{(x,y)\in I} F_t(x,y)} \quad (2)$$

,where I is the spatial position of the pixel. The size of human body is normalized by $F_t(x,y)$. Then, M is smoothed by averaging over a period of time to reduce noise. A key frame is detected at the local minimum of the smoothed M. So, the time span of the smoothed window has to be considered based on the velocity of actions. By using too small window, a lot of key frames are detected. On the other hand, using too large window will lead to miss detected key frames.

The number of frames in between the consecutive key frames, or the key frame interval as shown in Fig. 2, are not fixed based on the action movement, speed, and performer. So, the key frame interval for extracting features are automatically adapted by the (2) and are used in constructing history image.

B. Adaptive Motion History Image (AMHI)

To construct an adaptive key frame history image, we construct AMHI [20] within adaptive key frame interval as mentioned in section III-A. The history image is presented the layered silhouette within the interval, as shown in Fig. 2. The intensity is varied based on the time stamp within the key frame interval. The brighter intensity corresponds to the recent silhouette. AMHI represents motion direction and speed of the movement.

Given t_{pk} as a timestamp of the previous key frame and t_{ck} is a timestamp of the current key frame, $AMHI_t$ at time t is defined as in (3).

$$AMHI_t(x,y) = \begin{cases} t, if\ F_t(x,y) > 0 \\ 0, else\ if\ MHI_{t-1}(x,y) \le\ t_{pk}\ , for\ t > t_{pk} \\ MHI_{t-1}(x,y), otherwise \end{cases} \quad (3)$$

$AMHI_t$ is updated until t_{ck}. Then $AMHI_{t_{ck}}$ is used to construct feature descriptor in section III-D.

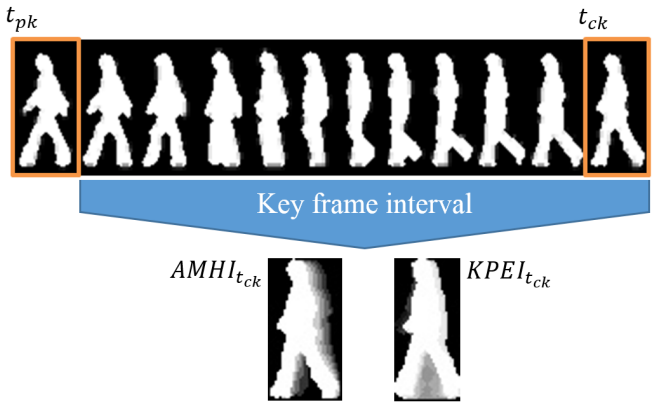

Fig. 2. KPEI and AMHI extracted from key frame interval of walking.

C. *Key Pose Energy Image (KPEI)*

KPEI construction is similar to AMHI as previously described in section III-B. KPEI is also constructed from the layered foreground within the key frame interval. Prior to constructing $KPEI_t$, the bounding box $R_t(i,j)$, which contains a blob from F_t, is segmented and aligned at the center of KPEI. Then the extracted foreground region locates at the center of an image $KPEI_t(i',j')$. Given (i_K, j_K) as the center of $KPEI_t$ and (i_R, j_R) as the center of R_t, $(i',j') = (i + (i_K - i_R), j + (j_K - j_R))$. $KPEI_t$ at time t is defined as in (4).

$$KPEI_t(i',j') = \begin{cases} KPEI_{t-1}(i',j') + t, \text{if } R_t(i',j') > 0 \\ 0, \text{else if } KPEI_{t-1}(i',j') \leq t_{pk} \\ KPEI_{t-1}(i',j'), \text{otherwise} \end{cases}, \text{for } t > t_{pk} \quad (4)$$

$KPEI_{t_{ck}}(i',j')$, which aligns the segmented and layered silhouettes within the key frame interval, indicates the body parts movement of the primitive action. Compared to $AMHI$, $KPEI$ does not considered in movement direction of the whole body. However, it focuses on the transformation of human body parts.

$AMHI$ and $KPEI$ at t_{ck} are used to create histograms of oriented gradient as an action description. If t is not the current key frame, both $AMHI$ and $KPEI$ are not used in feature process and are stored until the next key frame occurrence.

D. *Image Oriented Gradient*

At time t_{ck}, the histograms of image oriented gradient are constructed from $AMHI$ and $KPEI$. The gradient computation is based on Sobel operator in x and y dimensions, which are G_x and G_y, respectively. The direction of the gradient is defined as in (5).

$$\theta(x,y) = arctan\frac{G_y(x,y)}{G_x(x,y)} \quad (5)$$

Then, the local histograms of gradient are created for $AMHI$ and $KPEI$. The feature descriptor contains the concatenated local histograms of the history images.

IV. ACTION CLASSIFICATION

In the training process, feature descriptors which are extracted from the training set are used in dictionary learning. From [23], given training set $A = \{a_1, \dots, a_n\} \epsilon \mathbb{R}^{mxn}$ and a dictionary $D = \{d_1, \dots, d_p\} \epsilon \mathbb{R}^{mxp}$, the learning process is to find the p basis as in (6).

$$\min_D \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} \min_{\alpha^i} \left(\frac{1}{2}\|a_i - D\alpha_i\|_2^2 + \varphi(\alpha_i)\right) \quad (6)$$
$$\text{s.t. } \sum_{j=1}^{p} {d_j}^2 \leq 1$$

, where φ is a sparsity function and α is a sparse coefficient. The dictionary, which contains a set of basis from the training set, is used in the action classification.

For the testing process, a testing descriptor is decomposed and reconstructed by using Orthogonal Matching Pursuit (OMP) method. To classify action from the feature descriptor, the residual of the difference of the testing descriptor b and the sparse representation is considered by solving (7).

$$\min_{\alpha\epsilon\mathbb{R}^p} \|b - D\alpha\|_2^2 \quad (7)$$

In this work, each dictionary basis belongs to only one action. The residual of the sparse representation is used to classify action by finding the minimum residual from the dictionary basis.

At each key frame, the feature descriptor is decomposed and find the basis which has the minimum error. A class of each key frame can be identified by the basis label. The majority voting is used to classify action in each sequence.

V. EXPERIMENTAL RESULTS

The proposed framework system was tested with WEIZMANN human action dataset and robustness dataset [12]. Human action dataset contains actions of different performers which consist of walking, running, galloping sideways, jumping forward, jumping in place, jumping jack, bending, one hand waving, and two hands waving, as shown in Fig. 3. For robustness dataset, there are two sets of walking which contain the variation of viewpoint and deformation. The different viewpoints of walking vary from 0° to 45°, which increase 5° in each step. The deformations of walking in different scenarios are shown in Fig. 4. The sequences have the spatial resolution of 180x144 pixels. The datasets provided foreground mask extracted by using background subtraction.

In this work, 8 bins histogram of oriented gradient are created at the 3x3 local regions of AMHI and KPEI. In dictionary learning, the value of φ is 0.15. The precision of classification is computed by using leave one out cross validation. Actions are classified in each extracted feature representation within key frame interval and in each sequence.

Fig. 3. WEIZMANN dataset contains 9 actions, (a) walking, (b) running, (c) galloping sideways, (d) jumping forward, (e) jumping in place, (f) jumping jack, (g) bending, (h) one hand waving, (i) two hands waving.

Fig. 4 Robustness dataset contains 10 waling videos in the different scenarios, (a) walk with a dog, (b) swinging a bag, (c) walk in a skirt, (d) occluded feet, (e) occluded by a pole, (f) moonwalk, (g) limp walk, (h) walk with knee up, (i) walk with briefcase, and (j) normal walk.

For classification per key frame interval, the sparse representation is used to classify action as mentioned in section IV. The majority voting of the key frame labels are used to classify action in sequence.

The results of action classification on human action dataset in Table I indicated that the sparse representation of AMHI and KPEI descriptors have higher accuracy than using only AMHI or KPEI feature. The accuracy per key interval of KPEI is higher than AMHI. While, the accuracy of AMHI for a whole sequence is higher than KPEI. For some intervals, key frame interval is miss detected or the movement cycle is incomplete such as at the beginning and at the ending of the sequence. AMHI which focuses on the motion within the interval are effected with the problem of incomplete information of action cycle, such as half cycle. So the representation is misclassified for actions that have the similar motion pattern. While, KPEI which focuses on the transformation of body posture can handle the effect better than AMHI since the key posture is emphasize. However, we observed that by using only KPEI feature cannot classify actions which haves similar postures such as running and jumping forward, jumping jack and jumping in place, and two hand waving. By concatenating AMHI and KPEI features, the sparse representation has more information on both motion and posture. The results of the concatenated AMHI and KPEI descriptor have the highest classification rate in key frame and sequence classification.

For robustness dataset, the system used human action dataset as training data which consists of 9 actions. The viewpoint variation of walking sequences can be correctly classified from 0° to 25° for KPEI and from 0° to 30° for AMHI. The classification rate at each key frame interval dropped at 20°. Since, the training data contains only a viewpoint at 0°, increasing degree of viewpoints confused the classification. Walking at 45° is misclassified as jumping in place which have the performer facing a camera. For deformation variation, the results in Table II indicated that the sparse representation of oriented gradient can handle some deformation of walking scenarios. The results of walking with a dog and occluded legs indicated that AMHI cannot correctly classify actions which the significant body parts are occluded. While KPEI feature cannot classified action with occlusions such as occluded legs and occluded by a pole. Also, the limp walking which has different in both posture and motion from a normal walk led to misclassify by both features. The misclassification results included galloping sideway and jumping forward which share some similarities with walking.

By applying sparse representation on feature descriptors, the decompositions do not based on a training sample but also based on weight of several basis. So, the representation is more flexible and more suitable for a large data set.

VI. CONCLUSIONS

In this paper, the sparse representation of adaptive key frame interval descriptors is presented for classifying human actions. AMHI and KPEI are represented motion and posture pattern of human action within the key frame interval. The interval detection is based on the motion variation. The local histograms of oriented gradient of AMHI and KPEI are concatenated to construct feature descriptor. The training descriptors are used in dictionary learning to find basis for the sparse representation reconstruction. The classification is based on the basis which have the minimum residual in the reconstruction. The experimental results indicated that the proposed features can classify actions and has robustness in some scenarios.

TABLE I
COMPARISON OF CLASSIFICATION ACCURACY OF FEATURES ON WEIZMANN DATASET.

Feature	AMHI descriptor	KPEI descriptor	Concatenated AMHI and KPEI descriptor
Accuracy (%) (per key frame interval)	90.78	92.80	96.83
Accuracy (%) (sequence)	98.77	96.30	100

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY OF FEATURES ON IRREGULARITIES PERFORMANCE OF WALKING

Test sequence	AMHI descriptor	KPEI descriptor	Concatenated AMHI and KPEI descriptor
Walking with a dog	50	100	100
Swinging a bag	100	100	100
Walking in a skirt	100	100	100
Occluded Legs	25	50	50
Occluded by a "pole"	100	75	100
Sleepwalking	100	100	100
Limping man	66.67	50	66.67
Knees Up	71.43	71.43	71.43
Carrying briefcase	100	100	100
Normal walk	100	100	100

REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," JIVC, 28(6): 976-990, 2010.

[2] J. K. Aggarwal and M. S. Ryoo, "Human Activity Analysis: A Review," CSUR, 43(3), 16:1-43, 2011.

[3] R. Cutler and L. Davis, "Robust Real-Time Periodic Motion Detection, Analysis, and Application," PAMI, 22(8): 781-796, 2000.

[4] C. Thurau and V. Halaváč, "Pose primitive based human action recognition in videos or still images," In CVPR, 2008.

[5] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-evuelta, "Silhouette-based human action recognition using sequences of key poses," PR, 34(15), 1799-1807, 2013.

[6] A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing Action at a Distance," In ICCV, 2003.

[7] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of Oriented Optical Flow and Binet-Cauchy Kernels on Nonlinear Dynamical Systems for the Recognition of Human Actions," In CVPR, 2009.

[8] G. R. Bradski and J.W. Davis, "Motion segmentation and pose recognition with motion history gradients," Mach. Vision Appl., 13(3):174-184, 2002.

[9] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, "Behavior recognition via sparse spatio-temporal features," In VSPETS, 2005.

[10] Z. Zhang and D. Tao, "Slow Feature Analysis for Human Action Recognition," PAMI, 34(3), 436-450, 2012.

[11] N. T. Nguyen, H. H. Bui, S. Venkatesh, and G. West, "Recognising and monitoring high-Level behaviours in complex spatial environments," In CVPR, 2003.

[12] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. "Actions as Space-Time Shapes," In ICCV, 2005

[13] J. C. Niebles and L. Fei-Fei, "A Hierarchical Model of Shape and Appearance for Human Action Classification," In CVPR, 2007.

[14] N. Ikizler and P. Duygulu, "Histogram of oriented rectangles: A new pose descriptor for Human action recognition," JIVC, 27(10): 1515–1526, 2009.

[15] Z. Zhao and A. Elgammal, "Information theoretic key frame selection for action recognition," In BMVA, 2008.

[16] S. Baysal, M. CanKurt, and P. Duygulu, "Recognizing human actions Using Key Poses," In ICPR, 2010.

[17] M. Raptis and L. Sigal, "Poselet key-framing: a Model for human activity recognition," In CVPR, 2013.

[18] S. Hu, Y. Chen, H. Wang, and Y. Zuo, "Human action recognition based on spatial-temporal descriptors using key poses," Proc. SPIE 9301, International Symposium on Optoelectronic Technology and Application 2014: Image Processing and Pattern Recognition, 2014.

[19] K. Schindler and L. V. Gool, "Action snippets: How many frames does human action recognition require?," In CVPR, 2008.

[20] K. Lertniphonphan, S. Aramvith, and T. H. Chalidabhongse, "Feature Extraction for Human Action Classification using Adaptive Key Frame Interval," In APSIPA, 2014.

[21] X. Zhang, Y. Yang, L. C. Jiao, and F. Dong, "Manifold-constrained coding and sparse representation for human action recognition," PR, 46(7), 1819-1831, 2013.

[22] H. Cheng, Z. Liu, L. Yang, and X. Chen, "Sparse representation and learning in visual recognition: Theory and applications," SP, 93, 1408-1425, 2013.

[23] J. Mairal, F. Bach and J. Ponce, "Sparse Modeling for Image and Vision Processing," Foundations and Trends® in Computer Graphics and Vision, vol. 8, no. 2-3, 85-283, 2014.