A New Method of View-Invariant Human Activity Recognition

Han Su[1,2]
1 Key Lab of Virtual Reality and Visual Computing of Sichuan
Sichuan Normal University
Sichuan, Chengdu 610066
2 College of Computer Science
Sichuan Normal University

Wenjie Wang[2,3]
2 College of Computer Science
Sichuan Normal University
Sichuan, Chengdu 610101
3 Chengdu University
Sichuan, Chengdu 610106

Abstract—We present a new method to recognize human activity under uncertain view. The activity width image which is extracted from the sequence of silhouette is used to represent the activity features. It converts the spatial and temporal pattern to a grey-level image. View difference causes the deformation of activity width images and degradation of recognition accuracy. To solve this problem, we transform the feature images to the new images in the uniform view based on the activity templates using Iterative closet point (ICP). For reducing the complexity, we choose the center line and the contour of the activity width images to register and the affine transformation is adopted to transform. The similarity matrix is used to measure the difference between the transformed feature image and the templates. We evaluate our method on the CASIA database and the results show that our method is simple and efficient.

Keywords-human activity recognition; ICP; the activity width image; similarity matrix

I. INTRODUCTION

Human activity recognition is one of important and key parts in computer vision and understanding[1]. It applies in the field of Video surveillance, forensic analysis and so on. In initial stage, the researchers works in the lab condition and under many ideal assumption. They usually assume that the camera view angle is not change and the distance between the subjects and camera, the resolution is not bad.

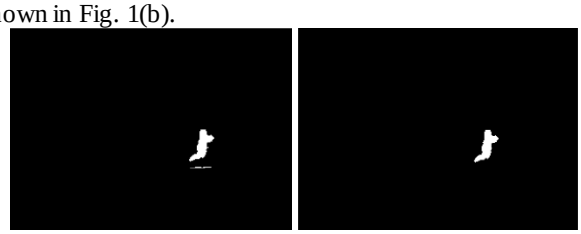
But in the reality environments, these assumption is impossible at all. Nowadays, many researchers carry on the view angle problem, and try to present the feature with view-invariant [2-8]. [2] proposes a self-similarity based descriptor for view-independent recognition and builds the Self-Similarity Matrix (SSM) for the action sequences. Different with other methods, they do not assume multi-view action samples and only compute distances between action representations for all pairs of time-frames and store them. Worapan Kusakunniran and et al. [3] present a novel method for gait recognition, they extract view-invariant features for cross-view gait. They use the input layer to normalize gait from arbitrary view and transformed onto the common view by domain transformation by invariant low rank textures, procrustes mean shape is used to present view-invariant features and procrustes distance is adopted to measure the similarities. Zheng[5] create a new dictionary learning framework for learning view-invariant spares representation, and model view shared features and a set of view-specific dictionaries. To extract the projective depth is proposed in [9].

In this paper, we also focus on the view angle problem. Unlike these methods, we try to correct the deformation by view angle and find the features which is view-invariant based on the features which is extract from the original sequence not the subject.

In section II, we segment the silhouette from each frame. The details of our method are explained in section III. And section IV and V are the experiments and conclusions.

II. PREPROCESSING

We consider that the background will change with the walking pedestrians, and the codebook method is used to model the background and segmentation. Codebook method [10] creates codebook for every pixel of the image. Each codebook includes one or more codewords. Every codeword is consisted of the maximum and minimum of thresholds and et al. In the modeling progress, the codebook of each pixel is matched when a new image is added to the sequence. If the incoming pixel is in the range of the detection threshold of codewords and satisfy some conditional requirements, it means the incoming pixel is matching with the codewords and the thresholds of codewords should be updated. If the incoming pixel is not matched to the codewords of codebooks, the new codeword will be created. In training processing, each pixel can correspond to one or more codewords. Fig. 1 shows the segmentation result by Codebook. Some holes are still in the image. To amend it, the morphological operators are used to amend it. The corrected image is shown in Fig. 1(b).



(a) (b)

Figure 1. The segmentation result by Codebook and morphological method.

III. FEATURE REPRESENT AND ANALYSIS

A. *The activity width image*

The width image is proposed in our earlier works[11,12]. We proposed the periodic sequence width image to present gait feature, the activity width image to describe temporal and structure features for human activity. The main idea is to convert the silhouette shape in each frame and the temporal shape changes over time to a single image. The image contains the structure and dynamic information at the same time. The effectiveness is proved in [11].The silhouette width is the basic and important element for the width image, so we name the image as the activity width image in this paper.

Fig. 2 is the activity width images. For a given human activity sequence, there are two steps to form the width images. Firstly, the silhouettes in the sequence is normalized the same size, and the width vector is calculated for each silhouettes. The width vector is composed by a series of silhouette width value. The width value is defined as the difference of vertical coordinate between the leftmost pixel and rightmost pixel on the edge in one row of frame. The all width values in the frame are the width vector. One frame is corresponding to one width vector, and the sequence is corresponding to a set of width vectors. The width vectors are sorted by the frame order in the sequence. Then, the width value is represented by gray value according to the position in the width vector and the order of width vector. Thus, the width vector set is converted to be a grey-level image. The image is the activity width image. The gray value present the width features in spatial and temporal space. Fig. 2(a) shows the width curve which is one width vector and contains the shape and static posture in freeze frame. Fig. 2 (b) and (c) are the width images from two different activities, (b) is running activity and (c) is crunch one.

B. *Contour registration*

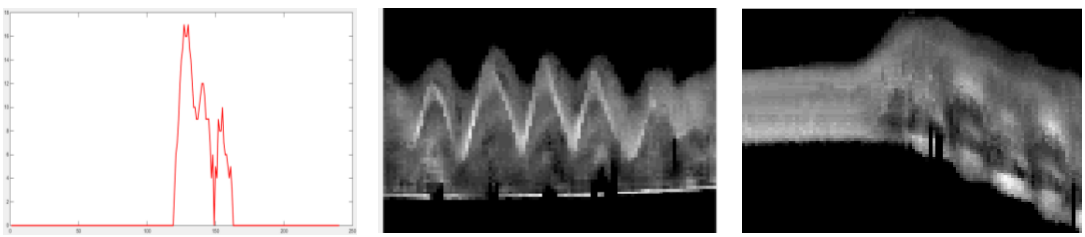
The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

In our earlier research[12], we evaluate the method on the CASIA database under the angle view and horizontal view, and the result is encouraging. But the result is built on the known conditions which are the view angle between the subjects and the camera position is known before test. The tests are consisted of two parts. One is in horizontal view, and the other is in angle view. The horizontal view sequence is compare with the other horizontal ones and the angle view sequence compare with the other angle view ones. In the reality environments, the view angle is not difficult to get and the subjects are not walking under certain view angle. It means that for one camera condition, the moving direction of subject is not always parallel or approximate parallel to the plane of projection, for multi-camera condition, to get the view angle is also a difficult problem. Usually, there are random view angles between the moving directions with the plane of projection. In our method, the range of view angle which we can deal with is from 0° to 60°. Because the over 60°, the body is self-occlusion seriously that cause the feature mismatch problem. The activity width image also faces that problem.

To be a view-invariant representation, we propose to transform the activity image with uncertain view angle to a new image under uniform view angle by the view transformation. It is consist of three steps. The first step is contour registration. The main aim is to find the closet angle and the most similar action between two activity images. To measure the similarity, ICP is an effective method which is adopted by many researches in different research field such as face recognition, gait recognition, image registration, reconstruct 2D or 3D surface, and point clouds registration and so on[13]. In our method, we adopt ICP to roughly match the view angle and find some candidate activities.

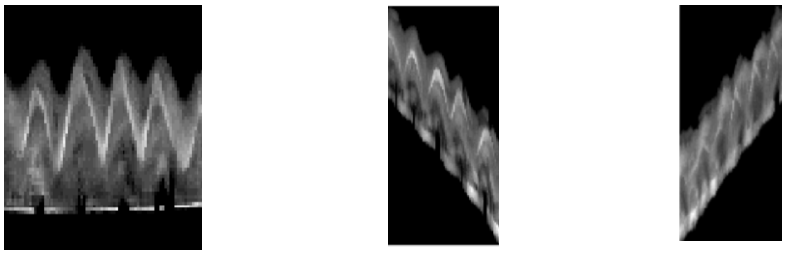
As we known, when the subject in 3D space shows in 2D image, it must be transformed by the projective transformation. The points on the subject will lose the depth information and be deformed in the process of the projective transformation. Owe to it, the activity width image which is calculated by the width of silhouette deforms at the same time. Especially, the positions of the corresponding pixels because of the changes of the scale, rotation, and et al. are changed However, the activity width image which preserve the structure information of subject and the temporal feature of action still keep mainly information after deformed by projection. The image contour and the center line of image are not changed strongly. In our method, we present to rough matching using ICP based on the image contour and the center line. It is shown in Fig. 3, (a) is in the horizontal view and (b),(c)are in uncertain view, (a) and (b) is the same action in different view. It is easy to find, the contours of the two images are similar to each other although the contour in Fig. 3(b) and(c) are deformed.

In training process, the activity templates are obtained. Each template is the average width images of the given action in horizontal view. For a given sequence under



(a) (b) (c)

Figure 2. The width curve and the activity width images.



(a) (b) (c)

Figure 3. The activity images in uncertain view.

uncertain view, the activity width image is calculated, and the contour of image is matching with all activity templates. The m nearest templates are the suspicious activity image.

The standard ICP which is presented by Besl and McKay[13] is a famous algorithm for rigid registration between two point sets. For each point in the source point set,ICP find the closet point in the reference set, iteratively revise to minimize the distance between the source point set and the reference one.

For a given activity width image, we define S{Si|Si∈ R^2,i=1,2,…,n} as the source point set, and the activity template G{Gi|Gi ∈ R^2,i=1,2,…,n} is the reference point set.

For each activity template, $f_i(R,t)$ is calculated.

$$f_j(R,t) = \sum_{i=1}^{n} \|G_i - (RS_i + t)\| \qquad (1)$$

where R and t are the rotation and Translation parameters. R and t are the two parts of transformation. The objective function is the minimum of $f_j(R,t)$.The optimal correspondence is

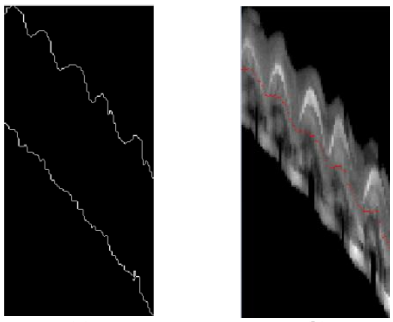
$$\min_{R,t,i\in\{1,2,\dots,n\}} \left(\sum \|G_i - (RS_i + t)\|_2^2\right) \qquad (2)$$
$$\text{s.t} \quad R^T R = I_n \,, \det(R) = 1$$

where $R \in \mathbf{R}^{n\times n}$, andt $\in \mathbf{R}^n$.The time complexity of this algorithm greatly grows when the point set is large. If all points of the activity image are calculated by (1), the time complexity is huge. To reduce the computing complexity, we choose the contour of image to represent the whole image. To speed up the point set matching process, we use the center line to initialize the initial value of ICP, and the points on the contour are adopted to iterate again. The center line is the main axis between the upper part contour and the lower part contour by the similar way of the SKICP[14]. The center line is shown in Fig. 4(b). The center lines are also obtained from the activity templates. The parameter k is the iteration index. The steps is as following,

- Step1, compute the distance between S and G to minimize ‖Gi-Si‖2.
- Step2,get the rotation transformation matrix R^k and the translation vector t^k and make $\sum \|G_i - (R^k S_i + t^k)\|_2^2 = min$
- Step 3, update the point set S, $S^{k+1} = \{S_i^{k+1} | S_i^{k+1} = R^k S_i^{\,k} + t^k\}$.
- Step 4, calculate the constraint $d^{k+1} = \frac{1}{n}\sum \left\|G_i^{k+1} - S_i^{k+1}\right\|_2^2$.
- If d^{k+1} is larger than the end constrain, then iterate to step 2, until d^{k+1} is less than the end constrain or satisfy the maximum iterations.

After get the initial rotation matrix and translation vector, the same steps are implemented between the contours of images. The rotation matrix Rj and translation vector tj are the transformation parameters for the given activity image and jth activity template. The number of activity template N is equal to the number of activity type.

For a given activity image, N pairs R and t is calculated using ICP and N new images are obtained by the front steps. Among N new images, m most similar activity templates according to the new images are treat as the undetermined candidates.



(a) (b)

Figure 4. The contour and the center line of the activity width image.

C. *View transformation*

The rotation matrix and translation vector which are only based on the center line and the contours is not enough to transform for view-invariant problem. The gray in the activity image embeds lots of features. The next step, the activity image is transformed by the affine transformation matrix to ensure the features are the independent of view. Although the main reason of deformation is the plane projection, in our case, to get the affine transformation matrix and transform the images to the horizontal view is right way. For each pixel(x,y) in the image, we use the normalized homogeneous coordinates. The pixel(x,y) is represented to (x,y,1).The affine transformation matrix is defined as

$$Af(x,y,1) = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \qquad (3)$$

where $\begin{bmatrix} a & d \\ b & e \end{bmatrix}$ is control the rotation, shear, scale, and flip, the vector t is (c f), $\begin{pmatrix} g \\ h \end{pmatrix}$ is the parameter of the

projection. $Af(x, y, 1)$ is easy to get by the matched points by R and t using ICP.

The given activity image is transformed to m fixed images by $Af(x, y, 1)$. There are m affine transformation matrixes according to the m most similar activity templates. The linear interpolation is used in our method.

D. Feature match

The activity width images are transformed in the same camera view after the processing of ICP and affine transformation. Each image is corresponding to m undetermined candidates. Feature match step is to find the most similar candidate. It is clearly, the contours are similar is not means the two images are belong to the same activity type and the gray and the variance are more important. The activity width image is regard as the feature image and texture image. To measure the similarity, the image similarity matrix[2] is used to feature analysis. The equation of image similarity matrix is same as the image self-similarity [16]. For the given activity width image I, I1,I2,…,Im are the transformed images. The activity templates are labeled as P1,P2,…,Pm, the image similarity matrix is denoted as S(Ii,Pi),

$$S(I_i, P_i) = \min_{dx,dy\in\{-1,+1\}} \sum_{(x,y)} |I_i(x+dx, y+dy) - P_i(x,y)| \quad (4)$$

where (x,y) is pixel position in both I_i and P_i, I_i (x,y) and P_i (x,y) are the gray value at (x,y) in I_i and P_i, respectively. The variance dx and dy are the range for a small search area. It means the similar point is in a 3*3 block in our method. Before computing the similarity matrix, the transformed images and template images should be normalized.

Using (4), the image similarity matrix set for I is obtained. Moreover, the histogram is created for each similarity matrix. The histogram represents the distributions of similarity between two images. The most similarity activity template is got through comparing the histograms using distance measure.

IV. EXPERIMENTS

To evaluate the performance of our method, we test it on the database of the Institute of Automation of the Chinese Academy of Sciences (CASIA). The database is in an outdoor database environment and the distance between the camera and subject is far. There are 1446 videos in this database. And the videos are taken in three different perspectives such as horizontal view, top down view, and angle view. The data are collected from 24 persons. The resolution of each frame is 320×240 and the resolution of subject in the frame is below 50×70 because the camera's distance. The activity types include walking, running, jumping, and bend, faint, crouch, wander and punching a car.

In this paper, we focus on the single activity of one subject not the multi-person interaction and view-invariant problem. We combine the horizontal view sequence and the angle view sequence into one dataset for testing the performance. The activity templates are obtained in training processing, and the number of candidate is 3. We choose the horizontal view to be the uniform view. It means that all activity images are transformed into the new feature images in the horizontal view. The activity include walking, running, jumping, and bend, faint, crouch are tested. The confusion matrix of our method is shown in Table 1.

For one sight, the result is not good as our earlier method. But the results of the earlier method is based on the certain view. The angle view was known in that test. In this experiments, the view angle is unknown and the sequence mixed into the set of the horizontal view through transformed the feature images. It is difficult than the known view condition.

TABLE I. CONFUSION MATRIX ON THE CASIA DATASET UNDER UNCERTAIN VIEW

	walk	**run**	**jump**	**bend**	**crouch**	**faint**
walk	0.8	0.03	0	0.17	0	0
run	0.04	0.79	0.17	0	0	0
jump	0	0.12	0.88	0	0	0
bend	0.12	0.01	0	0.84	0.03	0
crouch	0	0	0	0.07	0.72	0.21
faint	0	0	0	0.19	0.14	0.67

V. CONCLUSIONS

A new view-invariant activity recognition method is proposed in this paper. Firstly, Codebook algorithm is used to model the background and segment the silhouette images. To fix the holes and the disjoint silhouettes, the morphological operator are used to amend. Secondly, for each silhouette image, the width vector is calculated. The width curves which is corresponding to the width vectors are convert into the activity width image which is a grey-level image. The activity width image is the feature image of the silhouette sequence. All silhouette sequence under uncertain view angle are convert into feature images. The view angle problem causes the deformation of feature images. Then, ICP is used to register the unknown activity width image and the activity templates and find the candidates. To speed up the computing speed, we choose the center line to get the initial rotation matrix and translation vector and the contour of the feature image to image register. The affine transformation matrix based on the rotation matrix and translation vector is adopt to transform the images to be the same view angle. Finally, the image similarity matrix are computed to recognize the activities. In contrast to the other similar method, we directly transform the feature image not the subject. The results show that to transform the feature directly is a simple way to reduce the computing complexity and the width image is high-performance representation.

ACKNOWLEDGMENT

We would like to thank the Institute of Automatic, Chinese Academic of Science for providing the database. Our work is supported by the funding from the Key Project of Natural Science Foundation of Sichuan Provincial Education Department (Grant No.14ZA0029), Visual Computing and Virtual Reality Key Laboratory of Sichuan Province Foundation (Grant No.KJ201419) and 251 key talent project of Sichuan Normal University.

REFERENCES

[1] Pavan K. TuragaH, HRama ChellappaH, V. S. Subrahmanian H, Octavian Udrea, "Machine recognition of human activities: a survey," IEEE Trans. on Circuits and Systems for Video Technology, vol. 18, no. 11, pp. 1473–1488, 2008

[2] Imran N. Junejo,Emilie Dexter,Ivan Laptev,Patrick Pérez," Cross-View Action Recognition from Temporal Self-Similarities," Computer Vision ECCV 2008,Lecture Notes in Computer Science Volume 5303, 2008, pp 293-306

[3] Worapan Kusakunniran, Qiang Wu, Jian Zhang,Yi Ma,and Hongdong Li," A New View-Invariant Feature for Cross-View Gait Recognition," IEEE Transactions on Information Forensic Security, vol. 8, no. 10, 2013

[4] J. Zheng, Z. Jiang, J. Phillips, and R. Chellappa. Cross-view action recognition via a transferable dictionary pair. In BMVC, 2012

[5] Zheng, Z. Jiang, Learning View-invariant Sparse Representations for Cross-view Action RecognitionICCV2013

[6] V. Parameswaran and R. Chellappa. View invariance for human action recognition. IJCV, 2006

[7] R. Li and T. Zickler. Discriminative virtual views for cross-view action recognition. In CVPR, 2012

[8] J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In CVPR, 2011

[9] Nazim Ashraf, Chuan Sun b , Hassan Foroosh," View invariant action recognition using projective depth," Compuer Vision and Image Understanding,123,2014,pp.41-52

[10] Kyungnam Kim, Thanarat H. Chalidabhongse, David Harwood, Larry Davis, "Real-time foreground–background segmentation using codebook model," Real-Time Imaging, vol.11, no.3,2005,pp.172-185,doi: 10.1016/j.rti.2004.12.004

[11] Han Su , Guoyue Chen, "A new method of gait recognition independent of view angle," In Proc. Of ICMLC 2010, 3091–3096

[12] Han Su,Jiayun Zou," Human Activity Recognition Based On Silhouette Analysis Using Local Binary Patterns",FSKD 2013

[13] Paul J. Besl, Neil D. McKay. A method for registration of 3-D shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992. 14(2): 239-256.

[14] Ce Li, Xinying Luo, Shaoyi Du, Limei Xiao," A method of registration based on skeleton for 2-D shapes,"International Congress on 2012 5th Image and Signal Processing (CISP), DOI: 10.1109/CISP.2012.6469977

[15] R. Cutler and L. Davis, "Robust Real-Time Periodic Motion Detection, Analysis, and Applications," IEEE Trans. Pattern Recognition and Machine Intelligence, vol. 22, no. 8, pp. 781-796, Aug. 2000.

[16] Chiraz BenAbdelkade, Ross G Cutler, and Larry S Davis," Cross-View Action Recognition from Temporal Self-Similarities," EURASIP Journal on Advances in Signal Processing 2004, 2004:721765 doi:10.1155/S1110865704309236