Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **1**, January **2015**

A Semi-supervised Human Action Recognition Algorithm Based on Skeleton Feature

Hejin Yuan

Department of Computer
North China Electric Power University
619 Yonghua Street, Baoding, China
yhj_1977@163.com

Received October, 2013; revised August, 2014

Abstract. *A semi-supervised human action recognition algorithm using skeleton feature is put forward in this paper. In the method, the cumulative skeleton image and skeleton history image are firstly calculated as the feature representation of the human actions. Then, the label of unannotated actions is predicted through the constrained semi-supervised K-means clustering algorithm. Meanwhile the average cumulative and history skeleton images are generated as the model of each category actions. Finally, the nearest neighbour method is utilized to classify the observed action according to the correlation coefficients between its feature image and the pre-established templates. The experiments on Weizmann dataset demonstrate that our method is effective*
Keywords: Human action recognition, Cumulative skeleton image, Skeleton history image, Semi-supervised learning

1. **Introduction.** Human action recognition is very important in many computer vision applications, such as visual surveillance, intelligent perceptual interface, content based video retrieval. As a challenging issue, many considerable works have been done in recent literatures on this topic [1, 2, 3]. Yamato proposed a human action recognition method based on Hidden Markov Model [4]. In his method, the human images are divided into equant meshes and the pixels of each mesh are used as the feature vector for action recognition. Though, HMM is appropriate for human action modelling, it needs a lot of training samples and the self-adaptive determination of its structure and parameters is still very difficult. Bobick and Davis put forward a human action recognition algorithm with two temporal templates, named motion energy image and motion history image [5]. Weinland presented an action recognition approach using exemplar-based embedding [6]. In this method, the motion sequences are represented with respect to a set of discriminative static key-pose exemplars. Through the selected exemplars, the different sequences can be changed into same-length feature vector. And the recognition procedure can be greatly simplified by removing time related information, such as speed and length of an action. Though these methods' effectiveness has been verified through many different experiments, they all need a lot of labelled examples to train the recognition model. However, manually labelling much amount of human action sequences will be a time and labour consuming work. Moreover, the category of the action will change commonly with the variation of application environment and user requirements. So, the algorithms, which can achieve high accuracy with only a few labelled samples, are more favourable in practice. Since semi-supervised learning can combine the labelled and unlabelled data during

training to improve performance greatly, we applied semi-supervised learning for human action in [7] firstly. In ref [7], the features used for action recognition are motion energy image (MEI) and motion history image (MHI), proposed by Davis in [5]. In this paper, we propose another two new features, named cumulative skeleton image and skeleton history image, to represent human actions,.

The rest of this paper is organized as follows: The details about the proposed method, such as human action representation based on cumulative skeleton image and skeleton history image, measure between actions, unlabelled actions' class label prediction based on semi-supervised learning and nearest neighbour based classification, are given in section 2. In section 3, we evaluate our approach with well known Weizmann dataset before concluding in section 4

2. **Semi-supervised human action recognition based on skeleton feature.**

The main steps of our method include action representation, unlabelled actions' class label prediction based on semi-supervised learning and action classification with nearest neighbour method.

2.1. **Human Action Representation.**

Many different features can be used for recognition, such as colours, shapes, silhouettes and facial expressions [8]. The features used for human action recognition mainly can be classified into two categories: model-based and appearance-based. The appearance-based methods represent human action through lower level image features (such as silhouette, colour) and motion information (such as speed, optical flow and trajectory). In model analysis, model parameters are obtained from the image sequences through reconstruction. Though the model-based methods can provide much more useful information, the reconstruction procedure is neither robust nor reliable for real images since there are too noisy in these images. So, appearance-based features are commonly used in action recognition since they can be easily and robustly extracted from videos. The skeleton can be used for human action recognition when the human body can be truly segmented from the video background. Fig.1 shows the skeleton image of pjump and run action of lena in Weizmann dataset. From the figures, we can obviously find the skeleton features are very different in these actions. So, here we use the cumulative skeleton image (CSI) and skeleton history image (SHI) as the representation of human action.

The cumulative skeleton image is the accumulation of human skeleton in each frame. Let be a skeleton image sequence indicating human skeleton of motion. Here, skeleton image is a binary image, so in these images, the value of pixels on the skeletons are one, else are zero. The cumulative skeleton image is defined as:

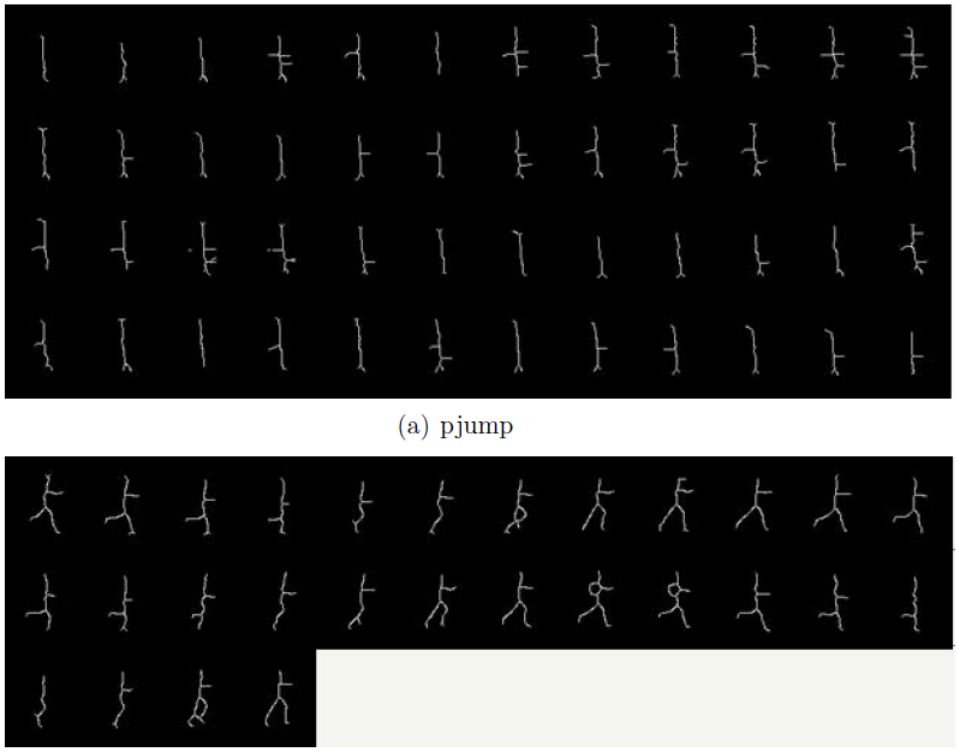
$$S_C(x, y, t) = \sum_{i=1}^{t} S(x, y, t) \tag{1}$$

The CSI of different actions of daria in Weizamman dataset is shown in Fig.2 (a).

The skeleton history image $S_H(x, y, t)$ is defines as:

$$S_H(x, y, t) = \begin{cases} 255 & \text{if} S(x, y, t) = 1 \\ \max(0, S_H(x, y, t-1) - 1 & \text{otherwise} \end{cases} \tag{2}$$

The SHI of different actions of daria in Weizamman data set is shown in Fig. 2 (b). Obviously, CSI shows the human skeleton and intensity of the actions in the image plane, while SHI reveals their temporal motion variation.



(a) pjump

FIGURE 1. Skeleton of action pjump and run of daria in Weizmann dataset

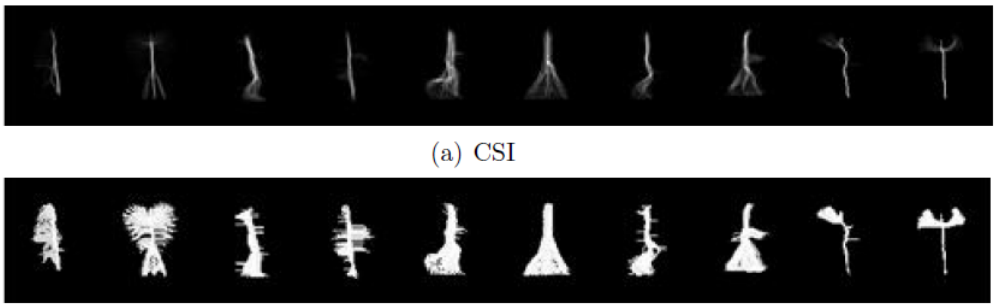
2.2. **Similarity Measure between Human Actions.** Bobick use Mahalanobis distance as the difference measure between the observed and the known actions. However, we find Mahalanobis distance is not very suitable for human action recognition since the scale of training dataset is not very large. So, in this paper, we use correlation coefficient, shown in formula (3), to measure the similarity between actions. Here,$c(i, j)$ is the correlation coefficients between actions i and j. S^i and S^j are the skeleton template image.

$$c(i,j) = \frac{S^i \cdot S^j}{\|S^i\| \, \|S^j\|} \tag{3}$$

Sorting the actions in Weizmann dataset according to their class label and then calculating the pair wise coefficient between different actions with CSI and SHI respectively. The result coefficient matrix of CSI and SHI is shown in Fig. 3. In the figure, the brighter the colour, the more similar between the actions.Obviously, our method can well distinguish out the actions since the coefficient between actions with same class label is much larger than others and the number of squares in the diagram is exactly equal to the number of action categories. Fig. 3 (b) also indicates that SHI is more robust than CSI for human action recognition since it contains temporal information of actions.

2.3. **Unlabelled data category prediction based on constrained semi-supervised K-means clustering.** In order to exploit the unlabelled data for action recognition, here we use the constrained semi-supervised K-means clustering algorithm proposed by Basu[9] to predict their categories. In this algorithm, the labelled data is used as the seed to initialize the K-means algorithm. And the cluster memberships of labelled data are kept unchanged during clustering. The category of unlabelled data is set identical to the labelled data in the same cluster.

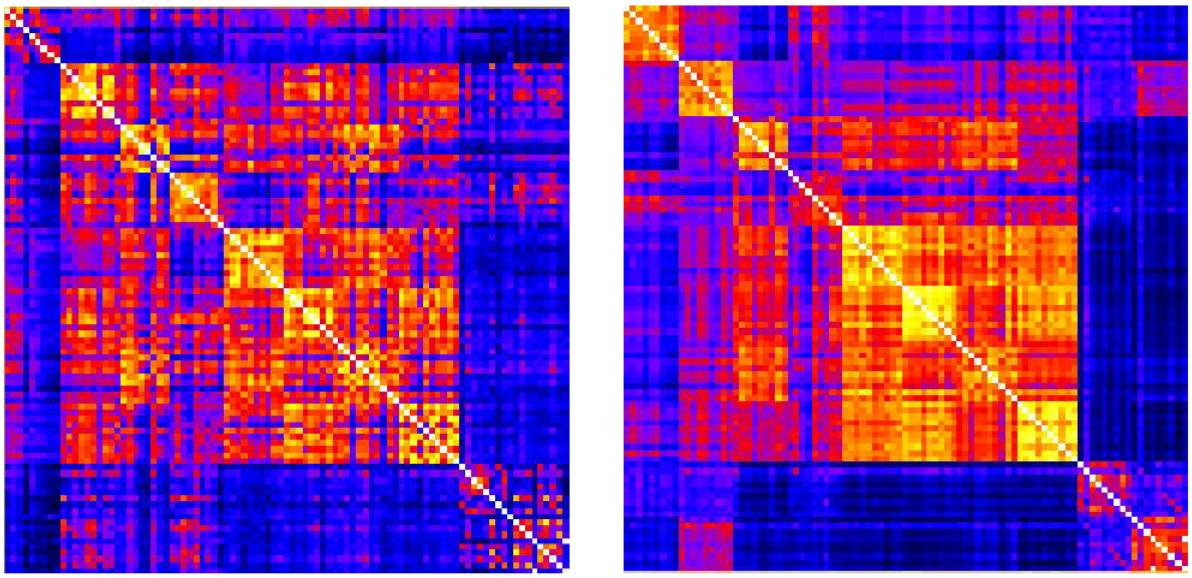
The detailed steps of constrained semi-supervised K-means algorithm are as follows:



(a) CSI

(b) SHI

FIGURE 2. CSI and SHI of daria in Weizmann dataset



(a) CSI (b) SHI

FIGURE 3. Correlation coefficients between actions in Weizmann dataset

Input: Set of data $X = \{x_1, x_2, ..., x_n\}$, number of clusters K, set $S = \bigcup_{l=1}^{K} S_l$ of labelled data.

Output: Disjoint K partitioning $\{X_l\}_{l=1}^{K}$ of X such that the K-means objective function is optimized.

Initialize:

$\mu_h^0 = \frac{1}{|S_h|} \sum_{x \in S_h} x$,for $h = 1, 2, ..., K$

$t = 0$

Repeat until convergence

a) For $x \in S$,if $x \in S_h$, assign x to the cluster h; else assign x to the cluster

$$h^* = \arg\max_h \frac{x \cdot \mu_h^{(t)}}{\|x\| \left\|\mu_h^{(t)}\right\|}$$

b)$\mu_h^{(t+1)} = \frac{1}{\left|X_h^{(t+1)}\right|} \sum_{X_h^{(t+1)}} x$

c)$t = t + 1$

2.4. **Action Recognition Based on Nearest Neighbour Method.** After predicting the categories of unlabelled training examples, we calculate the average CSI and SHI for each category action as follows:

$$\bar{S}_C^j(x,y,t) = \frac{1}{|X_j|}\sum\nolimits_{x_m \in X_j} S_C^m(x,y,t) \tag{4}$$

$$\bar{S}_H^j(x,y,t) = \frac{1}{|X_j|}\sum\nolimits_{x_m \in X_j} S_H^m(x,y,t) \tag{5}$$

Here, $\bar{S}_C^j(x,y,t)$ and $\bar{S}_H^j(x,y,t)$ are the average CSI and SHI of the jth category action respectively. And $\|X_j\|$ is the number of actions in training set with class label j.For the observed action x, firstly extracting its CSI and SHI with formula (1) and formula (2), denoted as S_C^x and S_H^x . Then, calculating their correlation coefficients to the average templates as follows:

$$Ce_x^j = \frac{S_C^x \cdot \bar{S}_C^j}{\|S_C^x\| \left\|\bar{S}_C^j\right\|} \tag{6}$$

$$Ch_x^j = \frac{S_H^x \cdot \bar{S}_H^j}{\|S_H^x\| \left\|\bar{S}_H^j\right\|} \tag{7}$$

The final category of the action , is determined by CSI and SHI, are respectively as follows according the nearest neighbour criteria:

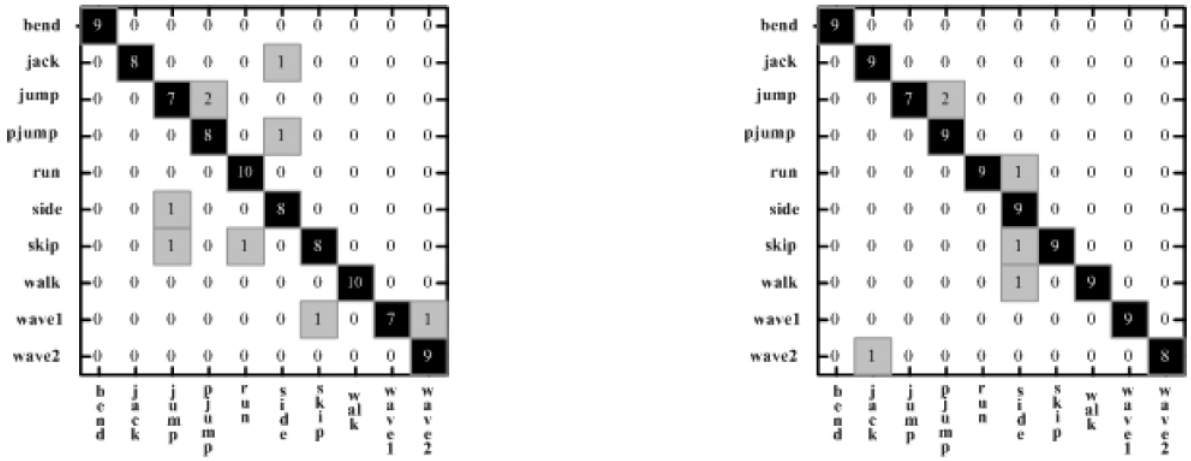
$$h_E^*(x) = \arg\max_j Ce_x^j \tag{8}$$

$$h_H^*(x) = \arg\max_j Ch_x^j \tag{9}$$

3. **Experiments and analysis.**

3.1. **Dataset.** For evaluating the proposed algorithm, this paper uses a publicly available dataset Weizmann, which is recently widely used in human action recognition algorithm evaluation. This dataset includes 10 natural actions: bending (bend), jumping jack (jack), jumping forward on two legs (jump), jumping in place on two legs (pjump), running (run), galloping-sideways (side), skipping (skip), walking (walk), waving one hand (wave1) and waving two hands (wave2), performed by 9 actors. The dataset in our experiments contains 10 actions and 93 videos, among them, the actor named lena, has two run, skip and walk action videos. Silhouettes extracted from backgrounds and original image sequences are also provided in the dataset. In our experiments, all recognition rates were computed with the leave-one out cross validation. Details are as follows: 8 out of the 9 actors in the database are used as the training samples and the 9th is used for evaluation. This procedure is repeated for all 9 actors and the rates are averaged.

3.2. **Processing.** We directly use the silhouettes provided in the dataset for subsequent processing. The CSI and SHI were firstly calculated for each action. For each training set, randomly select some actions and set their class label as unknown. Then, the constrained semi-supervised clustering algorithm is used to predict the class label of unlabelled data. After this procedure, the average templates for each action are calculated.



(a) CSI (b) SHI

FIGURE 4. Action confusion in classification

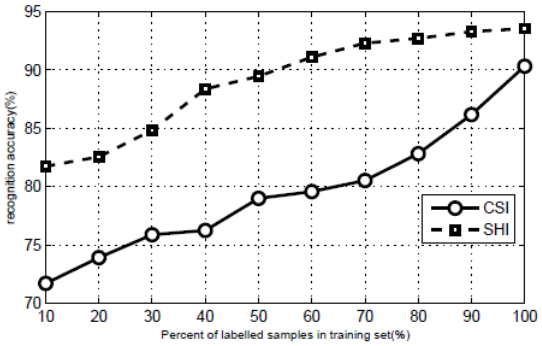
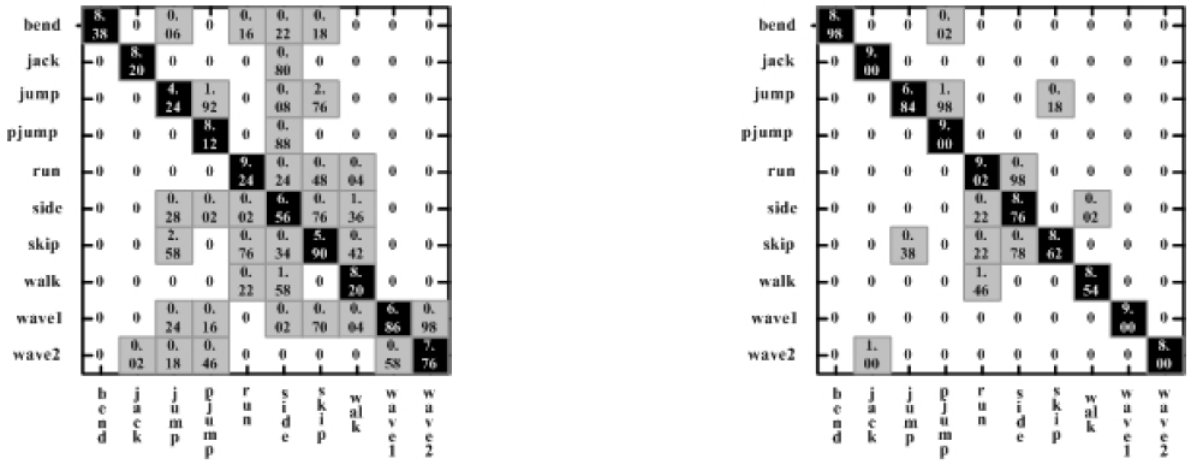


FIGURE 5. Average recognition rates vs. percent of labelled data

3.3. **Analysis.** The average recognition rates of our method for CSI and SHI are 90.32% and 93.55% when all the training examples are labelled. The confusion matrixes of CSI and SHI are also show in Fig.4. Obviously, the actions between jump and pjump, skip and run are easily to be wrong classified since they are very similar.

In comparison, the recognition rate of exemplar-based embedding method reported by Weinland[6] is 97.7% for 50 exemplars. The work of Ali et al. uses a motion representation based on chaotic invariants and reports 92.6% [10], while Wang and Suter[11] reported a recognition rate of 97.78% with an approach that uses kernel-PCA for dimensional reduction and factorial conditional random fields to model motion dynamics. The accuracy of our previous works in [7] is 95.70 and 93.55 respectively for MEI and MHI. So, the accuracy of our method is very close to those of state-of-art approaches. Comparing to the existed approaches, our method is much easier to be implemented and has less parameters to be adjusted.

For each category action, randomly selecting some sequence from training set and set their class label as unknown; then, using our method to experiment. The average recognition rate vs. percent of labelled examples is shown in Fig. 5. Here, the recognition rate is the average of 50 independent experiments. Obviously, the recognition rates rise gradually with the increase of percent of labelled examples. This result demonstrates that our approach is an effective semi-supervised human action recognition method. Its accuracy can reach above 80% even when the number of labelled data is very small (less than



(a) CSI (b) SHI

FIGURE 6. Action confusion in classification when the labelled training examples percent is 50%

10%), and the accuracy of our algorithm can reach above 90% when half of the training examples are labelled.

To examine and analyze which action sequences are incorrectly classified, we specifically show the confusion matrixes in Fig. 6 when the percent of labelled data is 50%. Obviously, the misclassified actions, marked with grey colour in confusion matrix, are consistent to their indistinguishability as shown in Fig. 3. The results in Fig.5 and Fig.6 also indicate SHI is more robust than CSI since it contains temporal information of actions.

Compare to our previous in [7], two new actions representation methods are proposed, named cumulative skeleton image and skeleton history image. Though the accuracy of our method in this paper is slightly lower than MEI and MHI, we think CSI and SHI are also suitable for human action recognition.

4. **Conclusion.** In this paper, we propose two new representation methods for human action recognition: cumulative skeleton image and skeleton history image. Then, the constrained semi-supervised K-means clustering algorithm is utilized to accommodate the challenge of obtaining accurate and detailed annotations of training data. Comparing to the existing methods, our approach can achieve high accuracy when only very small number of labelled training examples can be acquired. This is very important in practice. However, there are still many problems remains open, such as valuations on larger and realistic database, the variations of camera orientation and consideration of action semantics.

Acknowledgment. The work reported in this paper was supported by "the Fundamental Research Funds for the Central Universities (2014MS129)". We also acknowledge the anonymous reviewers for comments that lead to clarification of the paper.

REFERENCES

[1] Y.T. Du, F. Chen, W.L. Xu, et al, A suvery on the vision-based human motion recognition, *Acta Electronica Sinca*, vol. 35, no.1, pp. 84-90, 2007.

[2] J. Gu, X. Ding, and S. Wang, A Survey of activity analysis algorithms, *Journal of Image and Graphics*, vol. 14, no. 3, pp. 377-387, 2009.

[3] R. Popple , A survey on vision-based human action recognition. *Image and Vision Computing*, vol. 28, no. 6, pp. 976-990, 2010.

[4] J. Yamato, J. Ohya, and K. Ishii, Recognizing human action in time-sequential images using hidden markov model. *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 379-385, 1992.
[5] A. F. Bobick, and J. W. Davis, The recognition of human movement using temporal templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 3, pp.257-267, 2001.
[6] D. Weinland, and E. Boyer, Action recognition using exemplar-based embedding. *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp.1-7, 2008.
[7] H. Yuan, and C. Wang, Human action recognition algorithm based on semi-supervised kmeans clustering. *Trans. on edutainment VI*, pp. 227-236, 2011.
[8] K. Stelios, Statistical analysis of human facial expressions.*Journal of Information Hiding and Multimedia Signal Processing*, vol. 1, no. 3, pp. 241-260, 2010.
[9] S. Basu, A.Banerjee, and R. Mooney, Semi-supervised clustering by seeding. *Proc. of International Conference on Machine Learning*, pp. 19-26, 2002.
[10] S. Ali, A. Basharat, and M. Shah . Chaotic invariants for human action recognition. *Proc. of International Conference on Computer Vision*, pp. 1-8, 2007.
[11] L. Wang, and D. Sute, Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical mode. *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1-7, 2000.