21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Human Action Recognition Using Action Trait Code

Shih-Yao Lin*[,1], Chuen-Kai Shie[1], Shen-Chi Chen[2], Ming-Sui Lee[1,2],
Yi-Ping Hung[1,2]
[1]*Graduate Institute of Networking and Multimedia, National Taiwan University*
[2]*Dept.of Computer Science and Information Engineering, National Taiwan Univ.*
*{*d00944001, hung}@csie.ntu.edu.tw*

Abstract

Recognizing actions having similar movements is a challenging problem. Human action understanding task is divided into two issues in this paper. One is a classical action recognition task where we employ a probabilistic model to learn and recognize human actions. The other is action categorization task where we classify actions based on quantized human movement. An approach called ***Action Trait Code*** *(**ATC**) for human action classification is proposed to represent an action with a set of velocity types derived by the averages velocity of each body part. An effective graph model based on ATC classification is employed for learning and recognizing human actions. To examine recognition accuracy, we evaluate our approach on* ***Cornell Kinect Activity Database*** *and compare with a hierarchical maximum entropy Markov model (MEMM). Besides, the results on self-collected action database demonstrate that the proposed approach not only successfully achieves high recognition accuracy but also performs in real-time.*

1. Introduction

Human action recognition and categorization have been attracted much attention over the past few decades due to its applicability to many areas, including human computer interaction, game design, etc. However, the recognition task contains lots of challenging problems. This paper focuses on recognizing actions with similar movements and reduces the computational cost when applied to a huge dataset. According to a comprehensive survey [1], many previous studies on action recognition concentrated on using 2D videos [2] or still images [5]. However, those approaches are limited to express lateral motions only. Recently, 3D body joint locations have been widely used in human action recognition task [4] because it provides more explicit information for describing a human movements. Even if the input is 3D data, the state of arts may confuse with actions having similar poses.

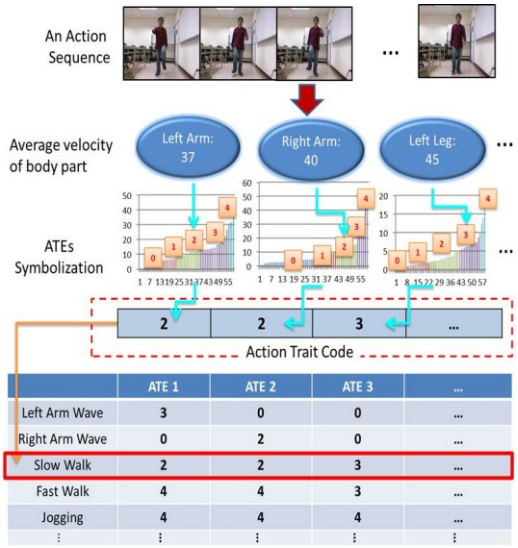


Fig. 1. Flowchart of Action Trait Code Extraction

The recognition task is performed by Microsoft Kinect sensor and Windows SDK for Kinect. Our system contains two major parts. One is human action classification which classifies an action by a set of velocity types derived by each body part's movement. The other is human action recognition task where we employ a probabilistic model based on classification task to learn and recognize the pose sequence of each human action. In the classification process, we propose an approach called Action Trait Code (Fig. 1) which uses the average velocity of body parts to yield a code describing the actions. We divide a human body into several body parts, such as left arm, right leg, etc. The average velocity of each body part in an action sequence is labeled as an Action Trait Elements (ATE). The fusion of each ATE can be encoded as

an ATC. Then we employ a graphical model based on ATC, a modified action graph [2][3], for learning and recognizing human actions. The experimental results on self-collected action database demonstrate that the proposed approach successfully delivers high recognition accuracy and is applicable to real-time applications. Moreover, we also test our approach on **Cornell Kinect Activity Database** [6] and compare with hierarchical maximum entropy Markov model (MEMM). The Cornell Kinect Activity Database contains twelve daily activities performed by four different people. The results have shown that our method achieves higher recognition accuracy than the two-layered MEMM.

The main contributions of this paper are enumerated as follows:

- An ATC classification is proposed to effectively divide tremendous action dataset into several smaller action dataset for increasing recognition accuracy and reduce the computational cost.
- The proposed graphical model based on ATC increases the recognition accuracy.

3. Action Trait Code

3.1 Action Trait Element

An Action Trait Element represents the average velocity of a specific body part using the mean distance of corresponding joints of 3D skeleton in a pose sequence (a human action). We divide a human body into N body parts. Let $\Psi = \{\varphi_1, \varphi_2, ..., \varphi_N\}$ be a velocity detector set, where φ_i denotes the i^{th} velocity detector. Given a body part J^i with a set of corresponding L joints, $\{j_1^i, ..., j_L^i\}$. The average velocity of J^i can be obtained by

$$\varphi_i(J^i) = \sum_{t=0}^{T-1} \sum_{k=1}^{L} dist(j_{k,t}^i, j_{k,t-1}^i) / T \qquad (1)$$

where T denotes the length of input action sequence and *dist*(.) is a distance function which is formulated with Euclidean distance.

A velocity discriminator is employed to quantify each body part's average velocity. If the size of action training data is q, the velocity discriminator uses k-means algorithm to classify the q average velocity values into k clusters. In other words, the range of ATE is from 0 to k. Fig. 2 illustrates the velocity discriminator construction process. Therefore, each average velocity of is tagged with a number as a symbol for Action Trait Element.

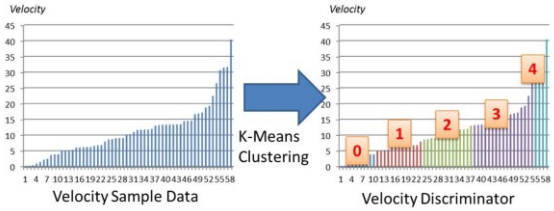


Fig. 2. Velocity Discriminator Generation

3.2 ATC Encoding based on ATC.

An ATC is like a specific body movement describer of an action. An ATC constructed by a combination of ATEs is symbolized from 0 to k. Each action can be encoded to an individual ATC. Moreover, ATC classifies same actions with fast and slow movements.

If we use an ATC of length N (using N body part) and each ATE is divided into k levels, then ATC codebook represents k^N action types. Thus, the action retrieval process using ATC classification reduces k^N times of computational cost. Furthermore, in order to prevent the over-fitting problem, we employ the k-means algorithm to cluster similar ATCs by calculating the L2 distance between two ATCs.

4. Action Recognition based on ATC

With the increasing action classes in our database, to measure the similarity between observed human actions and the huge collected action classes will cost dramatically. In our system, actions are pre-categorized by ATCs. In other words, a huge action database is grouped into several small action groups. Thus, the recognition process reduces lots of computational burden by measuring the similarity in a small part of action database only.

4.1 Action Classification

Our action classification task contains two steps. First, we calculate an ATC from the input action sequence. Second, we search the specific action dataset with matched ATC. However, because ATC is encoded using each body part's movement, various kinds of actions might be categorized in the same database. Therefore, a graphical model is adopted for the recognition process.

4.2 Graphical Model based on ATC

The Action Graph [2][3] is an useful graphic model approach to represent a dynamic human motion with a set of salient poses. The salient poses are shared among various actions. The proposed modified action graph G_{c_i} from the database of action

class c_i that encodes L actions with M salient postures $\Omega_{c_i} = \{\omega_{c_i,1}, \omega_{c_i,2}, ..., \omega_{c_i,M}\}$ can be represented as:

$$G_{c_i} = \{\Omega_{c_i}, A_{c_i,1}.., A_{c_i,L}\} \quad (2)$$

where each pose represents as a node; $A_{c_i,l} = \{p(\omega_{c_i,j} \mid \omega_{c_i,i}, \lambda_{c_i,l})\}_{i,j=1:M}^{l=1:L}$ denotes the transitional probability matrix of the l^{th} action $\lambda_{c_i,l}$ in an action dataset c_i.

According to the graphical interpretation, the recognition system is described as a quadruplet:

$$\Phi_{c_i} = (\Lambda_{c_i}, \Omega_{c_i}, \Upsilon_{c_i}, G_{c_i}) \quad (3)$$

where $\Upsilon = \{ p(x_{c_i} \mid \omega_{c_i,1}),\ p(x_{c_i} \mid \omega_{c_i,2}), .. ,\ p(x_{c_i} \mid \omega_{c_i,M}) \}$. The action modeling process, Φ_{c_i} involves three major steps: (1) extract salient postures Ω_{c_i} from training data, (2) model each posture by likelihood functions Υ, and (3) construct the action graph G_{c_i}. Each salient pose is a set of similar poses, which is obtained by clustering the training poses. We cluster these poses into M salient postures by K-means algorithm. We assume that the distribution of the salient postures can be approximated statistically independently. The posture model can be represented as the joint distribution of points.

$$p(p_{\omega_{c_i}} \mid \omega_{c_i}) = \prod_{i=1}^{n} \sum_{j=1}^{c} \pi_{j,\omega_{c_i}} N(p_i \mid \mu_{j,\omega_{c_i}}, \Sigma_{j,\omega_{c_i}}) \quad (4)$$

where $N(.)$ is a Gaussian function; In recognition task, we present a modified action graph based on ATC to recognize human actions. The action recognition process is to seek the most likely action λ^* from a set of action models $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_L\}$. Let $X = \{x_1, x_2, ..., x_T\}$ be a set of pose sequence derived from the input action sequence with T frames. The action recognition process can be formulated as:

$$\begin{aligned}
\lambda^* &= \arg\max_{\lambda} p(c_i, X, \lambda) \\
&= \arg \max_{\lambda_{c_i} \in \Lambda_{c_i}, S_{c_i} \subset \Omega_{c_i}} p(X_{c_i}, S_{c_i}, \lambda_{c_i} \mid c_i) p(c_i) \\
&= \arg \max_{\lambda_{c_i} \in \Lambda_{c_i}, S_{c_i} \subset \Omega_{c_i}} p(x_{c_i,1}, ..., x_{c_i,n} \mid s_{c_i,1}, ..., s_{c_i,n}, \lambda_{c_i}, c_i) \\
&\qquad \times p(s_{c_i,1}, ..., s_{c_i,n} \mid \lambda_{c_i}, c_i) p(\lambda_{c_i}, c_i)
\end{aligned} \quad (5)$$

where $S_{c_i} = \{s_{c_i,1}, ..., s_{c_i,n}\}$ represents the corresponding posture sequence derived from X; $p(c_i)$ is a prior based on the confidence of ATC classification $p(c_i)=1$. Assume that $p(c_i)=1$ and $x_{c_i,t}$ statistically depends only on $s_{c_i,t}$. $x_{c_i,t}$ is statistically independent of λ_{c_i} given by S_{c_i}, and $s_{c_i,t}$ only depends on its previous state $s_{c_i,t-1}$. Thus, we can reformulate Eq. (5) as:

$$\begin{aligned}
\lambda^* = \arg \max_{\lambda_{c_i} \in \Lambda_{c_i}, S_{c_i} \subset \Omega_{c_i}} & \prod_{t=1}^{N} p(x_{c_i,t} \mid s_{c_i,t}, c_i) \times p(\lambda_{c_i}, c_i) \\
& \times p(s_{c_i,1}, ..., s_{c_i,n} \mid \lambda_{c_i}, c_i)
\end{aligned} \quad (6)$$

where $p(x_{c_i,t} \mid s_{c_i,t}, c_i)$ expresses the probability of observation, $x_{c_i,t}$, derived from salient posture $s_{c_i,t}$.

5. Experimental Results

We evaluate our approach on two datasets: self-collected database and Cornell Activity database [6]. In the experimental setting, the proposed approach divides human body into four parts: left upper limbs, right upper limbs, left lower limbs, and right lower limbs.

Evaluate on Self-collected dataset. The self-collected dataset contains 20 different actions performed by four people. The proposed action database contains many actions with similar pose sequences. In this experiment, we compare the recognition accuracy between the proposed approach and Action Graph. The Action graph approach trains the joint locations of the full body. The confusion matrix of Action Graph is shown in Fig. 3(a). Since Action Graph measures similarity by full body pose sequence, the Action Graph misunderstands the actions with similar poses sequences (e.g. clap and play table tennis). Fig. 3(b) shows the confusion matrix of the proposed approach. As we can see from the results, the proposed approach based on ATC classification solves the problem by estimating body's movements separately.

Evaluate on Cornell Kinect Activities Dataset. The Cornell Kinect Activity Database is performed by 12 different activities performed by four different people in different indoor environments. We compare our proposed method against three models, naïve classifier based on SVM, one-layer MEMM, and a hierarchical maximum entropy Markov model (MEMM). We experimented with "have seen" setting [4], which halved testing subject's data and include one half in the training data set. Table 1 shows the precision and recall scores for each approach. As shown in Table 1, the proposed method achieves higher recognition accuracy (precision/recall of 97.7/97.2) than other approaches. Moreover, our method is able to recognize actions from similar movements, such as "rinsing mouth",

"brushing teeth", and "wearing contact lens." To access experimental results, please visit the project webpage:
http://csie.ntu.edu.tw/~d97944010/research/icpr12/

6. Conclusion

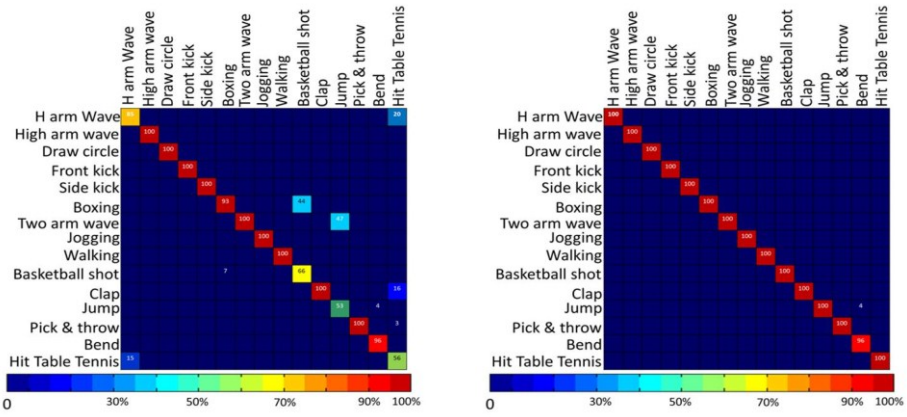
In this paper, a new approach, called Action Trait Code, for human action classification is proposed. We also present a graph model based on ATC for human action recognition. To evaluate our approach, two datasets are tested: self-collected dataset and Cornell Kinect Activity dataset. Experimental results demonstrate that our approach successfully increase recognition accuracy.

Acknowledgements

This work was supported in part by the National Science Council, Taiwan, under grants NSC 100-2622-E-002-005-CC2 and NSC 101-2622-E-002-005-CC2.

References

[1] J. K. Aggarwal, M. S. Ryoo, " Human Activity Analysis: A Review," ACM Computing Survey, 2011.
[2] W. Li., Z. Zhang, Z. Liu, "Expandable Data-Driven Graphical Modeling of Human Actions Based on Salient Postures." TCSVT 18(11):1499–1510, 2008.
[3] W. Li., Z. Zhang, Z. Liu, "Action Recognition Based on A Bag of 3D Points," CVPRW, 2010.
[4] J. Sung, C. Ponce, B. Selman, A. Saxena, "Unstructured Human Activity Detection from RGBD Images," ICRA, 2012.
[5] B. Yao, X. Jiang, A. Khosla, "Classifying Actions and Measuring Action Similarity by Modeling the Mutual Context of Objects and Human Poses," ICML, 2011
[6] Cornell Kinect Activity dataset: http://pr.cs.cornell.edu/humanactivities/



(a) (b)

Fig. 3. Confusion matrixes: (a) Confusion matrix for Action Graph; (b) confusion matrix for our approach

Table 1. Experimental result of Naïve classifier, MEMM model, Hierarchical HEMM model and our approach.

Location	Activity	Naïve Classifier		One-layered MEMM		Hierarchical MEMM		Our approach	
		Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
Bathroom	rinsing mouth	73.3	49.7	70.7	53.1	64.1	70.9	100	100
	brushing teeth	81.5	65.1	81.5	75.6	96.7	77.1	100	100
	wearing contact lens	87.8	71.9	87.8	71.9	79.2	94.7	100	100
	Average	80.9	62.2	80.0	66.9	79.1	80.9	**100**	**100**
Bedroom	talking on the phone	70.2	67.2	70.2	69.0	88.7	90.8	100	75.0
	drinking water	64.1	31.6	64.1	39.6	83.3	81.7	80.0	100
	opening pill container	48.7	52.3	48.7	54.8	93.3	77.4	100	100
	Average	61.0	50.4	61.0	54.5	88.4	83.3	**93.3**	**91.6**
Kitchen	cooking (chopping)	78.9	28.4	78.9	29.0	70.3	29.0	100	100
	cooking (stirring)	44.6	45.8	44.6	45.8	74.3	45.8	100	100
	drink water	52.2	51.5	52.2	51.5	88.8	52.4	100	100
	open pill container	17.9	62.4	17.9	62.4	91.0	62.4	100	100
	Average	48.4	47.2	48.4	47.4	81.1	74.3	**100**	**100**
Overall Average		**64.3**	**53.2**	**63.1**	**56.2**	**82.8**	**79.5**	**97.7**	**97.2**