Hybrid Generative-Discriminative Recognition of Human Action in 3D Joint Space

Zhe Wu
Department of Automation
Shanghai Jiao Tong University
Shanghai, China, 200240
zjuwuzz@gmail.com

Xiong Li
Department of Automation
Shanghai Jiao Tong University
Shanghai, China, 200240
lixiong@sjtu.edu.cn

Xu Zhao
Department of Automation
Shanghai Jiao Tong University
Shanghai, China, 200240
zhaoxuhong@gmail.com

Yuncai Liu
Department of Automation
Shanghai Jiao Tong University
Shanghai, China, 200240
whomliu@sjtu.edu.cn

ABSTRACT

We propose a novel human action recognition method based on the generative feature mapping over *3D human body joint sequences*. The proposed method relies on Hidden Markov Model (HMM), but differs from the previous methods in the way of incorporating HMM and discriminative classifier, aiming to capture more discriminative information. Firstly, we use HMMs to model the joint sequences of human body. Then the *Posterior Divergence* is used to build feature mappings from the trained HMMs. The derived feature mappings map a variable-length joint sequence to a fixed-dimension feature vector which will be delivered to SVM for classification. We evaluate the proposed method and related methods on a large number of 3D joint sequences. The experimental results show its competitive performance, in comparison with other state-of-the-art methods.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Motion*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*

General Terms

Algorithms, Design, Experimentation

Keywords

Action Recognition, Posterior Divergence, HMM

1. INTRODUCTION

Human action recognition is an important yet challenging topic in computer vision field [3, 5]. In this paper, we present an approach to solve this problem based on *3D human body joint position time-sequences*. The 3D joints data are obtained from the Motion Capture system or the video based pose tracking system, where at each time point, the 3D positions of a group of specified joints are recorded. Because the variance of viewpoint and illumination has been removed, 3D joint sequence provides a robust representation of human action, in comparison with video based 2D representation. However, the high dimensionality of the 3D joint space is still a critical restriction to achieve computational effectiveness. Besides, it remains a great challenge to extract discriminative features from 3D joints data.

Hidden Markov Model [6] and its variants have been successfully used in 3D joint-sequence based action recognition due to their ability in modeling variable-length sequences. In [7], spatial-temporal HMM (ST-HMM) is proposed to model the pair-wise spatial dependency of connected joints. ST-HMM, using maximum a posterior (MAP) classifier, outperforms HMM (also use MAP classifier) as it exploits the additional spatial information. Although HMM and its variants are good at modeling the variable-length motion sequences and exploiting hidden information of the 3D joints data, their classification ability working with MAP classifiers is not as powerful as discriminative classifiers in most situations [1]. Despite its ability in classification, discriminative models, e.g. Support Vector Machine (SVM), are lack of the ability to accommodate variable-length sequences.

HMM-boost [2] is a method designed to combine the advantages of the two methods. In HMM-boost, the joints are grouped into 141 overlapping *parts* (from holistic to local), where each part is a combination of some related joints. For each part of each action class, a HMM is trained ($\#class \times 141$ in total). These HMMs forms 141 MAP classifiers which are used as weak classifiers and combined by the multi-class Adaboost to make a comprehensive decision. Although the recognition accuracy is improved, training a large number of HMMs is time-consuming. Further, its performance is sensitive to the construction of parts.

Aiming at capturing discriminative information efficiently, we propose to combine HMM with discriminative classifiers based on Posterior Divergence (PD) [1], which is different from ST-HMM [7] or HMM-boost [2]. PD is a principled method to derive explicit feature mapping from generative

MM'12 October 29 - November 2, 2012, Nara, Japan.

models, which captures discriminative information of samples by fully exploiting hidden variables. The dimension of the derived feature mappings and subsequently the feature vectors is fixed, so they can be delivered to classifier directly. In addition, we would show that PD can be more powerful in multi-class classification through an extension (referred to as PD-3). Compared with HMM-boost, our method only need a much smaller number of features to get the better performance.

In sum, the contributions of this paper are two-fold: (1) we propose a hybrid generative-discriminative method for human action recognition based on 3D joint sequences. As far as we know, there is little work in action recognition using such kind of methods. (2) we propose an extension for PD which could be more robust in multi-class classification. Experiments on a large amount of 3D joints data show that our method outperforms HMM and its variants.

2. POSTERIOR DIVERGENCE

We first revisit Posterior Divergence (PD) [1] in this section. PD is a hybrid generative-discriminative scheme deriving feature mappings or score spaces from generative models. It is derived from the log likelihood of the incremental EM algorithm [4] which only looks at one or a few samples to update the model in each iteration.

Let $\mathcal{X} = \{\mathbf{x}^1, ..., \mathbf{x}^N\}$ be the set of samples; $H = \{h_1, ...h_M\}$ be the set of hidden random variables; i indexes samples and m indexes hidden variables. Let $\mathbf{x}^c$ be the sample to be looked at the t-th iteration; $\mathcal{X}_{+c} = \mathcal{X} \cup \{\mathbf{x}^c\}$. Let $P(\mathbf{x}\,|\,\theta)$ be the model estimated from the set $\mathcal{X}$; $\{Q^i\}_i$ be the approximations of posterior distributions $\{P(H\,|\,\mathbf{x}^i, \theta)\}_i$ where $\mathbf{x}^i \in \mathcal{X}$. Similarly, let $P(\mathbf{x}\,|\,\theta_{+c})$ be the model estimated from the set $\mathcal{X}_{+c}$; $\{Q^i_{+c}\}_i$ be the approximations of posterior distributions $\{P(H\,|\,\mathbf{x}^i, \theta_{+c})\}_i$ where $\mathbf{x}^i \in \mathcal{X}_{+c}$.

In the t-th iteration of the incremental EM algorithm, the approximate posterior distribution $Q^{c,t}$ and the model θ are updated in E step and M step respectively. Then the contribution of $\mathbf{x}^c$ to the entire sample set could be written in the log likelihood:

$$\mathcal{L}(\mathbf{x}^c) = \sum_{i=1}^{N}[-\mathcal{F}(Q^i_{+c}, \theta)] - \sum_{i\neq c}^{N}[-\mathcal{F}(Q^i, \theta)] \qquad (1)$$

where $\mathcal{F}$ denotes the variational free energy [1]. $Q(H)$ can be factorized according to samples due to the i.i.d assumption. Then $\mathcal{F}$ can be written as:

$$\mathcal{F}(Q, \theta) = \sum_i \mathcal{F}(Q^i, \theta) = \sum_i \mathrm{E}_{Q^i}\left[\log \frac{Q^i(H)}{P(\mathbf{x}, H\,|\,\theta)}\right] \qquad (2)$$

Without loss of generality, we use the following factorizations: $Q(H) = \prod_m Q(h_m)$; $P(\mathbf{x}, H\,|\,\theta) = P(\mathbf{x}\,|\,\mathrm{pa}_x, \theta) \cdot \prod_m P(h_m\,|\,\mathrm{pa}_m, \theta)$ where pa_x and pa_m are the parent variable sets of $\mathbf{x}$ and h_m respectively. Substitute Eq. (2) into Eq. (1) and apply the above factorizations, we can get the log likelihood in the form of entropy or cross entropy. Further, note that Q^i and Q^i_{+c} are approximately equal when $|\mathcal{X}|$ is relative large, i.e., $Q^i(h_m) \approx Q^i_{+c}(h_m)$. By applying this approximation, we have:

$$\begin{aligned}\mathcal{L} \approx\ & \underbrace{[\sum_{i\neq c}^{N} \mathrm{E}_{Q^i} \log \frac{P(\mathbf{x}|\mathrm{pa}_x, \theta_{+c})}{P(\mathbf{x}|\mathrm{pa}_x, \theta)}}_{\phi^{\mathbf{x}}_{pd}} + \underbrace{\mathrm{E}_{Q^c} \log P(\mathbf{x}|\mathrm{pa}_x, \theta_{+c})]}_{\phi^{\mathbf{x}}_{fit}} \\ & + \underbrace{[\sum_{i\neq c}^{N} \mathrm{E}_{Q^i} \log \frac{P(h_1|\mathrm{pa}_1, \theta_{+c})}{P(h_1|\mathrm{pa}_1, \theta)}}_{\phi^{h_1}_{pd}} + \underbrace{\mathrm{E}_{Q^c} \log P(h_1|\mathrm{pa}_1, \theta_{+c})}_{\phi^{h_1}_{fit}} \\ & - \underbrace{\mathrm{E}_{Q^c} \log Q^c(h_1)]}_{\phi^{h_1}_{ent}} + \underbrace{...}_{h_2...h_{M-1}} + [\underbrace{...}_{\phi^{h_M}_{pd}} + \underbrace{...}_{\phi^{h_M}_{fit}} - \underbrace{...}_{\phi^{h_M}_{ent}}] \end{aligned} \qquad (3)$$

where $\phi^v_{pd}(\mathbf{x}^c)$ denotes posterior divergence function which measures how much $\mathbf{x}$ affects the model; $\phi^v_{fit}(\mathbf{x}^c)$ denotes the fitness function which measures how well the sample fits the model; $\phi^v_{ent}(\mathbf{x}^c)$ denotes entropy function which measures how uncertain the fitting is. Thus for a generative model and input $\mathbf{x}$, we have a set of feature mappings:

$$\Phi^c(\mathbf{x}^c) = \left(\phi^{\mathbf{x}}_{pd}, \phi^{\mathbf{x}}_{fit}; \phi^{h_1}_{pd}, \phi^{h_1}_{fit}, \phi^{h_1}_{ent}; ...; \phi^{h_M}_{pd}, \phi^{h_M}_{fit}, \phi^{h_M}_{ent}\right)^T \qquad (4)$$

The obtained feature mappings share the same number of dimensions and could work with discriminative classifier like SVM straightforwardly. As shown in [1], working with zero-one loss linear classifier, the performance of PD is proved to be at least as good as that of MAP classifier.

2.1 Practical Strategies

Posterior Divergence, deriving feature mappings from generative models, provides a way to exploit hidden information. In real applications, two typical strategies are adopted. *The first strategy* trains a generative model θ for the training samples of all classes. The PD feature mapping for each sample $\phi^c = ({\phi^c_{pd}}^T, {\phi^c_{fit}}^T, {\phi^c_{ent}}^T)^T$ is obtained over the trained model θ (Eq. (4)). We refer to this strategy as PD-1. *The second strategy* trains S class-conditional generative models $\{\theta_1, \theta_2, ..., \theta_S\}$ for S classes respectively. The obtained feature mapping is a concatenation of those over all class-conditional models,

$$\phi^c = ({\phi^c_1}^T, \cdots, {\phi^c_S}^T)^T$$

which is referred to as PD-2. Compared with PD-1, PD-2 seeks to capture the class-conditional information.

The third strategy is an extension of PD-2. PD-2 simply concatenates the feature mappings over all class-conditional models $\{\theta_1, \cdots, \theta_S\}$, which actually assumes that all models can represent the sample well and are equally important to the sample $\mathbf{x}^c$. This assumption not always holds as a sample from the class s_1 (or model θ_{s_1}) can be an outlier of the class s_2 (or model θ_{s_2}) when $s_1 \neq s_2$. As a substitution, we measure how well a model θ_s represent the sample $\mathbf{x}^c$, and use this measure to weight the feature mapping over that model. The likelihood of a sample over a model θ_s, i.e., $P(\mathbf{x}^c\,|\,\theta_s)$, is a natural measure or weight for such purpose. This weight prefers those models fitting or explaining $\mathbf{x}^c$ well. The overall feature mappings can be expressed as,

$$\phi^c = ({\phi^c_1}^T P(\mathbf{x}^c\,|\,\theta_1), \cdots, {\phi^c_S}^T P(\mathbf{x}^c\,|\,\theta_S))^T \qquad (5)$$

This strategy is referred to as PD-3 throughout this paper and illustrated in the right part of Fig. 1.

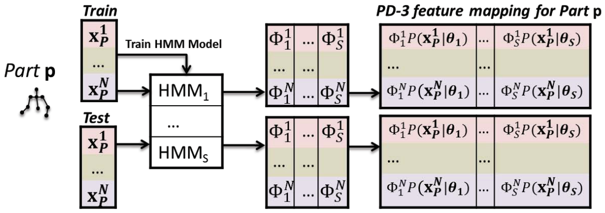
Train
Part p
Train HMM Model
HMM$_1$
...
HMM$_S$
Test
PD-3 feature mapping for Part p
$\mathbf{x}_P^1$
$\mathbf{x}_P^N$
Φ_1^1 ... Φ_S^1
Φ_1^N ... Φ_S^N
$\Phi_1^1 P(\mathbf{x}_P^1|\boldsymbol{\theta}_1)$ ... $\Phi_S^1 P(\mathbf{x}_P^1|\boldsymbol{\theta}_S)$
$\Phi_1^N P(\mathbf{x}_P^N|\boldsymbol{\theta}_1)$ ... $\Phi_S^N P(\mathbf{x}_P^N|\boldsymbol{\theta}_S)$

Figure 1: The illustration of extracting feature mappings the p-th part. $N=$#sequences, $S=$#classes.

3. HUMAN 3D ACTION RECOGNITION

In this section we apply PD to human 3D action recognition. As shown in [2, 7], exploiting spatial context among joints is useful to recognition. Like [2], we decompose all joints into several overlapped parts (groups). In each part, the spatial context would be exploited. However, distinct with [2], we only need few holistic parts instead of 141 parts (see Fig. 3). We derive the feature mappings for each part separately and finally combine the feature mappings of all parts. To make the representation of human pose invariant to the size, initial orientation and absolute position of body, we normalize the joints positions using the method in [2].

For each action class and each part, we train a HMM. Let random binary vector $q_{1\times M}^t$ and $y_{1\times N}^t$ denote the hidden state (M states) and output state (N states). The parameter set of a HMM is denoted by $\theta = \{\pi, A, B\}$, where π is the initial state probability, A is the transition probability and B is the output probability. Using Baum-Welch algorithm [6], we can compute the model parameter θ. Let τ and G respectively be the initial state probability and the transition probability of the approximate posterior. The free energy function (Eq. (2)) of HMM can be expressed as:

$$\mathcal{F}(Q,\theta) = \mathrm{E}_Q\left[\sum_{i=1}^{M} q_i^0 \log\frac{\tau_i}{\pi_i} + \sum_{t=0}^{T_c-1}\sum_{i,j=1}^{M} q_i^t q_j^{t+1}\log\frac{G_{ij}}{A_{ij}} - \sum_{t=0}^{T_c}\sum_{i,j=1}^{M,N} q_i^t y_j^t \log B_{ij}\right] \tag{6}$$

Based on the model θ, we estimate the approximate posteriors $\{Q^i(q_i^0, q_i q_j, q_i y_j \,|\, \tau, G)\}_i$ through maximizing the variational lower bound $-\mathcal{F}(Q,\theta)$ with respect to Q^i, and then estimate θ_{+c}. Then the feature mappings can be derived through Eq. (3). For example:

$$\phi^c = \mathrm{vec}\left(\left\{\sum_{k\neq c} g_{ij}^k \log\frac{a_{ij,+c}}{a_{ij}},\ g_{ij}^c \log a_{ij,+c},\ g_{ij}^c \log g_{ij}^c\right\}_{i,j=1}^{M,M} \cup \left\{\sum_{k\neq c} h_{ij}^k \log\frac{b_{ij,+c}}{b_{ij}},\ h_{ij}^c \log b_{ij}\right\}_{i,j=1}^{M,N}\right) \tag{7}$$

Similarity, for each part $p \in \{1,\cdots,P\}$ and each action class $s \in \{1,\cdots,S\}$, we train the model θ_{sp} and estimate the above quantities needed by PD, and compute its feature mapping ϕ_{sp}^c (Eq. (7)). Then, for each part, we combine the feature mappings over all classes using Eq. (5). Finally, we concatenate the feature mappings for all parts. The final feature mapping for a given action sample $\mathbf{x}^c$ is:

$$\Phi^c : \mathbf{x}^c \rightarrow \left(\underbrace{{\phi_{11}^c}^T, ..., {\phi_{1S}^c}^T}_{part\,1:\,\mathbf{x}_1^c}, ..., \underbrace{{\phi_{P1}^c}^T, ..., {\phi_{PS}^c}^T}_{part\,P:\,\mathbf{x}_P^c}\right)^T \tag{8}$$

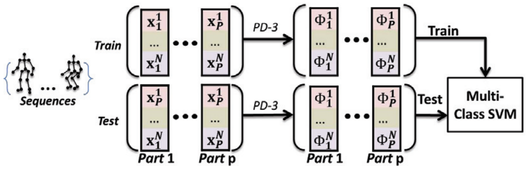

Figure 2: The framework of our method which combines feature mappings of all parts. P denotes parts.

The complete framework of our method is shown in Fig. 2. First, we train HMMs for each part of each action class; second, compute the feature vectors of all joint sequences using the feature mappings derived by PD (and PD-3) and deliver to multi-class SVM for classification.

4. EXPERIMENTS

To validate the proposed 3D action recognition algorithm, we test it on a large 3D MoCap dataset which is obtained from the CMU MoCap database and has been used by [2, 7]. Our dataset consists of 8 classes of actions: *walk, jog, jump, handshake, punch, climb ladders, sit and wave.* We manually collected 515 3D joints sequences (video clips), each of which contains a action loop with about 90 frames. The number of samples of the above 8 actions are 114, 74, 75, 17, 60, 17, 56, 70 respectively. Throughout of the experiment, the number of hidden states is set to $K = 3$.

In the *first experiment*, we compare the three strategies in our framework using 2 holistic body parts. We run the experiment 10 rounds. In each round, we randomly select 50% samples from each action class for training and the remainder for test. The recognition accuracies of PD-1, PD-2 and PD-3 (proposed) are respectively 90.88%, 96.78% and 99.10%. As stated in Section 2.1, the effectiveness of PD-3 can be seen by a surprisingly high recognition accuracy, even with only two parts. The confusion matrix of PD-3 is shown in Table. 1

In the *second experiment*, we evaluate the robustness of our method with respect to the numbers of parts, in comparison with 4 related methods: (1) 2Gram with SVM, (2) HMM with MAP classifier, (3) HMM-Boost [2], (4) ST-HMM with MAP classifier [7]. 2Gram is actually the transition probability matrix of the observed states of a sequences, and serves as the baseline method for sequences recognition. Nine numbers of parts from 2 to 41 are tested. The criterion of parts selection is similar with [2], i.e., from holistic parts (e.g., coordinates of all arm related joints) to local parts (e.g., coordinates of 3 connected joints). Fig. 3 reports the average recognition accuracy of the tested methods. Since HMM and ST-HMM trains a HMM for all joints, treating the body as a whole part, they are denoted by two dotted lines. The recognition accuracy of different classes are reported in Table. 2.

The results show that the proposed hybrid method outperforms the 4 related methods, especially when the number of parts is small. This indicates that by using the proposed hybrid method, we only need few complimentary

Table 1: The average confusion matrix of PD-3 with 2 holistic parts over 10 rounds of tests.

%	*#sample*	**Walk**	**Run**	**Jump**	**Wave**	**Punch**	**Shake**	**Sit**	**Climb**
Walk	114	**100**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Run	74	1.75	**98.25**	0.00	0.00	0.00	0.00	0.00	0.00
Jump	75	0.00	0.00	**99.12**	0.00	0.00	0.00	0.88	0.00
Wave	17	0.00	0.00	0.00	**100**	0.00	0.00	0.00	0.00
Punch	60	0.00	4.44	0.00	0.00	**95.56**	0.00	0.00	0.00
Shake	17	0.00	0.00	0.00	0.00	0.00	**100**	0.00	0.00
Sit	56	0.00	0.00	0.00	0.00	0.00	0.00	**100**	0.00
Climb	70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100**

Table 2: Recognition accuracy (%) of all actions

%	**Walk**	**Run**	**Jump**	**Wave**	**Punch**	**Shake**	**Sit**	**Climb**	**Overall**	Time(s)
(*#sample*)	114	74	75	17	60	17	56	70	512	
2Gram-SVM	94.72	91.05	93.68	100	92.67	82.22	69.29	96.00	90.97	2
HMM-MAP	93.06	100	97.37	100	91.33	97.78	75.71	94.86	93.28	62
HMM-Boost [2] ($P = 141$)	99.30	100	100	96.67	98.67	97.78	96.07	97.43	98.69	10023
ST-HMM[7]	98.61	98.68	98.68	94.44	83.33	66.67	87.50	91.43	93.44	1100
HMM-PD-1[1]($P = 26$)	100	100	97.37	91.11	100	91.11	97.86	99.43	98.69	825
HMM-PD-2 ($P = 11$)	100	100	99.12	96.30	95.56	100	100	98.10	98.97	860
HMM-PD-3 ($P = 6$)	100	100	100	100	96.67	100	100	99.05	99.49	453

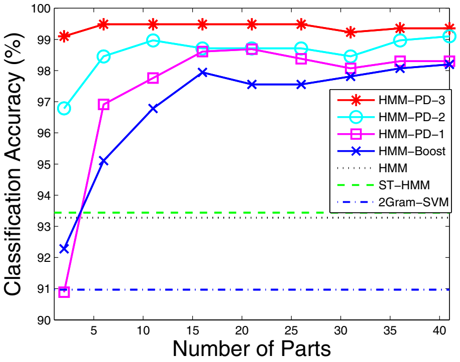

Figure 3: Evaluation of the robustness with respect to different number of parts.

parts instead of a large amount of holistic and local parts to fully capture the discriminative information for classification. Among the three strategies of PD, the proposed PD-3 achieved the highest recognition accuracy (99.4% with 6 parts), and is clearly more robust than PD-1 and PD-2. At the same time, the performance of HMM-Boost [2] is not satisfied when the number of parts is small, but increasing the number of parts would improve the performance. The performance of ST-HMM [7] is better than traditional HMM, but still not very satisfied due to the limitation of the MAP classifier. In multi-class classification, MAP classifier could easily fail in handling similar action sequences from different classes. While the feature mappings derived by our method could encode more discriminative information than simply likelihood and performed much better.

5. CONCLUSIONS

In this paper, we present a novel framework for human action recognition. We derive the discriminative feature mappings for 3D joint sequences, which could give an overall measure over different action classes and different body parts. Then we combine the traditional HMM with discriminative classifiers and enhance the classification performance. The experimental results show that our method outperforms other methods significantly. In particular, our method only need a few complimentary parts to reach the highest recognition rates.

6. ACKNOWLEDGEMENTS

This work is supported by China National 973 Program (2011CB302203).

7. REFERENCES

[1] X. Li, T. S. Lee, and Y. Liu. Hybrid generative-discriminative classification using posterior divergence. *CVPR*, pages 2713–2720, 2011.

[2] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. *ECCV*, pages 359–372, 2006.

[3] T. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, 2006.

[4] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, 89:355–368, 1998.

[5] R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108(1-2):4–18, 2007.

[6] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE.*, 77(2):257–286, 1989.

[7] Q. Zhao, L. Wang, H. H.S.Ip, and X. Zhou. Human 3d motion recognition based on spatial-temporal context of joints. *ICPR*, pages 2740–2743, 2010.