PLOS | ONE

RESEARCH ARTICLE

Jointly Learning Multiple Sequential Dynamics for Human Action Recognition

An-An Liu*, Yu-Ting Su, Wei-Zhi Nie*, Zhao-Xuan Yang

School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China

* liuanan@tju.edu.cn (AAL); weizhinie@tju.edu.cn (WZN)

Citation: Liu A-A, Su Y-T, Nie W-Z, Yang Z-X (2015) Jointly Learning Multiple Sequential Dynamics for Human Action Recognition. PLoS ONE 10(7): e0130884. doi:10.1371/journal.pone.0130884

Editor: Daoqiang Zhang, Nanjing University of Aeronautic and Astronautics, CHINA

Received: October 22, 2014

Accepted: May 26, 2015

Published: July 6, 2015

Data Availability Statement: All relevant data is available via the Harvard Dataverse (http://dx.doi.org/10.7910/DVN/YDDJ9F).

Funding: This work was supported in part by the National Natural Science Foundation of China 573 (61472275, 61170239), the Tianjin Research Program of Application Foundation, and 574 Advanced Technology (15JCYBJC16200), the grant of Elite Scholar Program of 575 Tianjin University (2014XRG-0046).

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Discovering visual dynamics during human actions is a challenging task for human action recognition. To deal with this problem, we theoretically propose the multi-task conditional random fields model and explore its application on human action recognition. For visual representation, we propose the part-induced spatiotemporal action unit sequence to represent each action sample with multiple partwise sequential feature subspaces. For model learning, we propose the multi-task conditional random fields (MTCRFs) model to discover the sequence-specific structure and the sequence-shared relationship. Specifically, the multi-chain graph structure and the corresponding probabilistic model are designed to represent the interaction among multiple part-induced action unit sequences. Moreover we propose the model learning and inference methods to discover temporal context within individual action unit sequence and the latent correlation among different body parts. Extensive experiments are implemented to demonstrate the superiority of the proposed method on two popular RGB human action datasets, KTH & TJU, and the depth dataset in MSR Daily Activity 3D.

Introduction

Human action recognition is an essential issue in computer vision and machine learning due to its wide and significant applications on multimedia content analysis and retrieval, human computer interaction and so on [1]. Recently, the importance is strongly highlighted by the urgent need for intelligent video surveillance in security sensitive environments.

Motivation and Overview

Although the current methods have shown the superior performance on this task, there still exist two problems.

- Only leveraging the global characteristics of one action while seldomly considering the body part information. One kind of methods utilize the bag-of-visual-word model with space-time interest point detectors and descriptors for global visual representation [2] [3]. The silhouette-based method is another representative method, which has been widely utilized for action recognition [4]. However, this kind of methods highly rely on accurate foreground

extractions, which seriously limits their application. Since human action usually shows a dynamic visual pattern, many sophisticated sequential modeling methods have been designed for this task. Grammar-based methods [5, 6] have been successfully applied for complex human activity recognition. Pei *et al.* [7] proposed a Stochastic Context Sensitive Grammar-based method for video event prediction. They constructed the hierarchical structure to represent the spatial and temporal relationships between the sub-events. The syntactic-based methods are another widely used method to represent human activity with high level temporal logic complexity. Hamid *et al.* [8] proposed to encode the sequential dynamic within human activities with a set of local action units. Consequently, they handled the problem of discovering action dynamic by feature selection. Graphical model-based methods [9] have recently attracted great interests due to its strong ability for sequential modeling. Lv *et al.* [10] trained the hidden Markov model with 3D joint trajectories to model the dynamic within human motion. Han *et al.* [11] utilized the conditional random fields model for the identification of continuous human activity. As the extension of the conditional random fields model, multiple advanced graphical models, including the hidden conditional random fields [12], the latent dynamic conditional random fields [13], the bidirectional-integrated random fields model [14], the semi-Markov model [15], etc., were developed for sequential modeling. Although these powerful methods have shown great performances on this task, they only consider the sequential dynamics of entire body within one action. Since the visual features of different body parts might lie in different feature subspaces, it would be difficult to learn the latent state spaces of one action by simply considering the sequential visual features of the entire body. Consequently, it is mandatory to take advantage of the characteristics of the body and parts together.

- Ignoring the sequential dynamics of body part. Recently, much more work has been done on the part-based method since part-based representation and modeling is more discriminative by focusing on local regions [16]. The most famous part-based method is the deformable part models (DPM) [17] which can simultaneously learn the characteristics of local and global body regions and their correlation with the latent support vector machine. Motivated by the DPM method, Wang *et al.* [18] [19] leveraged the motion patterns of both part regions and body region to construct a graphical model with the latent states for human action recognition. To avoid these highly structured models, Sharma *et al.* proposed the expanded parts model to automatically discover the parts and learn corresponding discriminative templates with their respective locations from a pool of candidate pars. Tian *et al.* [20] further improved the 2D DPM method, which used to be utilized for the still image, and designed the 3D DPM method to model the local and global spatiotemporal action pattern. We also propose the part-regularized multi-task structural learning method for both multiple-view and single-view action recognition [21]. This method can couple the body-based classification and the part-based classification to benefit intrinsic relatedness sharing across multiple action categories and consequently augment the performance of action recognition. These methods ignore the temporal context of partwise features for action modeling. Therefore, the specific sequential modeling methods need to be developed to take advantage of the sequential partwise features for temporal dynamic modeling.

To tackle both problems, it is mandatory to develop a method, which can take advantage of both global temporal context and partwise temporal context within one action to discover both sequence-specific structure and the sequence-shared correlation for action modeling. In this paper, we propose the multi-task conditional random fields model and explore its application on the task of human action. First, we partition human body into several parts with the prior knowledge of body structure and propose the partwise spatio-temporal action unit sequence

(ST-AUS) to represent the temporal context of individual body part during one action. Second, we propose the multi-task conditional random fields (MTCRFs) model to jointly learn sequential dynamics of entire body and individual parts and the correlation inbetween. We demonstrate the superiority of the proposed method on on two RGB human action datasets, KTH [22] & TJU, and the depth dataset in MSR Daily Activity 3D [23].

Contributions

The main contributions of the proposed method can be summarized as follows.

- Motivated by the theory of multi-task learning (MTL) [24], we originally propose the multi-task conditional random fields (MTCRFs) model. Different from the direct feature-level fusion and decision-level fusion, the proposed MTCRFs can learn both temporal structure within individual partwise spatio-temporal action unit sequence and transfer the latent correlation inbetween. Therefore, it can preserve the dynamics of individual sequence while sharing the complementary information of different body parts.
- We propose the partwise spatio-temporal action unit sequence (ST-AUS) to represent the multi-level visual dynamics. Individual ST-AUS can represent the part-specific visual dynamic while multiple ST-AUSs together can convey the latent correlations among multiple body parts. Therefore, the proposed ST-AUS representation can well depict the diverse visual characteristics of each action video.
- We contribute a novel human action dataset (TJU) to the community. TJU contains 22 types of human actions. There are totally 1760 action sequence. The synchronized RGB/depth/skeleton sequences of one action were taken with a Microsoft Kinect sensor with 20fps frame rate and 640 × 480 resolution. To obtain the satisfactory skeleton data, the dataset was recorded in the lab and no occlusion was set. However, it is still challenging since there are more complex and similar actions and both light and dark environments are concerned. This dataset can be downloaded from http://dx.doi.org/10.7910/DVN/YDDJ9F.

The rest of the paper is organized as follows. Section 2 briefly introduces the framework of the proposed method. Section 3 and 4 introduce the proposed spatio-temporal action unit sequence and the multi-task conditional random fields model, respectively. Section 5 explains the experimental method and Section 6 illustrates the experimental results. At last, we conclude the paper in Section 7.

Framework

The proposed method aims to automatically identify the action performing in one query video by discovering and modeling both partwise and bodywise dynamics. It contains two key steps, ST-AUS representation and MTCRFs modeling.

1) **ST-AUS representation**: We propose the partwise spatio-temporal action unit sequence (ST-AUS) as shown in Fig 1. We utilize the prior knowledge of body structure to define seven body parts, e.g., head, left/right limbs, left/right legs, and left/right feet. One specific part region lasting for T frames is considered as an action unit. Then the action units belonging to one part region in an action video is considered as a partwise action unit sequence. Each partwise action unit sequence focuses on the dynamic of specific body area. Consequently, each action video can be represented in multiple sequential feature spaces.

2) **MTCRFs modeling**: With the ST-AUS representation, we propose the multi-task conditional random fields (MTCRFs) model. The partwise spatio-temporal action unit sequences can be considered as the sequential observations of MTCRFs to discover the latent state space

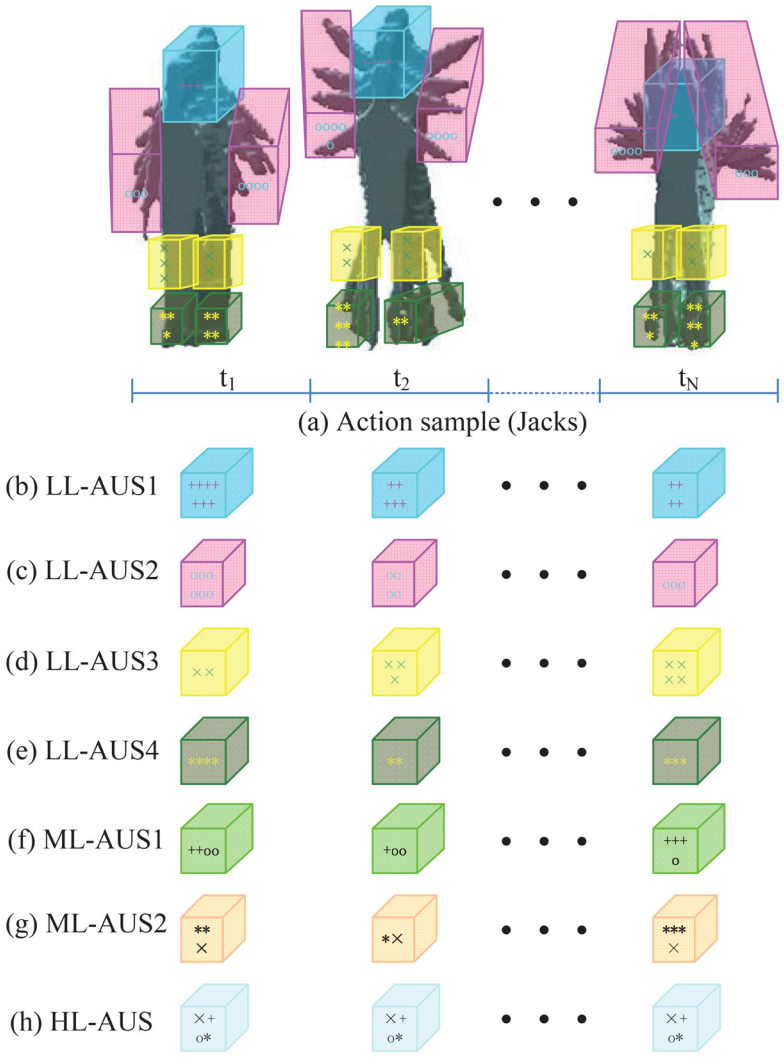


Fig 1. Partwise spatio-temporal action unit sequence. Note that *,×,+,o respectively denote the space-time interest points in different part areas. Different colors denote different body parts.

doi:10.1371/journal.pone.0130884.g001

and the transition inbetween. Since different partwise spatio-temporal action unit sequences might have different temporal dynamics, they cannot be linked or fused directly. In MTCRFs, each sequential observation is designed to connect to one sequential hidden state and all the hidden state sequences are correlated to the sequence label. Therefore, the proposed MTCRFs can flexibly learn the temporal dynamics of individual sequence and discover the correlation among them.

Spatio-temporal Action Unit Sequence

The proposed spatio-temporal action unit sequence(ST-AUS) representation contains three main steps (Fig 1).

First, we extract the local saliency descriptors. The extraction of local space-time feature can be accomplished by local saliency point detection and description. The popular local space-

time interest point detectors and descriptors [25] [26] [27] [28] [29] can be implemented for this step. With the extracted local space-time interest points, one video can be considered as a collection of local spatiotemporal points. We will further partition them into different groups depending of the prior knowledge of human body structure.

To achieve the body structure information, we implement two methods for body part localization: 1) For the classic RGB human action datasets, we implement the part model-based method [17] to localize 7 body parts (head, left/right limbs, left/right legs, and left/right feet). Fig 2a shows the samples from KTH. In each image, the big box denotes the localization of human body. The 7 small boxes denote the localized part regions. 2) For the recent datasets recorded by the Kinect sensor, the skeleton data can be directly used for body part localization. Fig 2b and 2c show the samples of the skeleton-based localization results on TJU and MDA.

With the localized part regions, we can obtain the centers of individual regions and then utilized them to group the space-time interest points into different part-induced categories. Depending on different feature pooling methods, the proposed ST-AUS can be classified into three levels.

Low Level (LL) ST-AUS: LL ST-AUS represents the basic partwise visual dynamics. The left and right limbs/legs/ feet are respectively considered as one category to avoid the sparse interesting points in individual regions. There are totally 4 kinds of parts, including head/limb/leg/ foot and consequently the space-time interest points can be grouped into the corresponding categories. Then, we can learn individual codebook with the K-means algorithm. The specific part regions in adjacent F frames is defined as a spatio-temporal action unit (ST-AU) which represent the local saliency of a specific body part locating in a special spatio-temporal volume. Therefore the dynamic evolution of a specific body part performing an action can be represented by the sequential ST-AUs, which is defined as the spatio-temporal action unit sequence (ST-AUS). With the learned dictionary corresponding to each part, the standard Bag-of-Words (BoW) scheme can be implemented on each ST-AU for BoW representation and consequently ST-AUS can be represented by the sequential BoW features. Fig 1b-1e show the BoW representations for four different partwise ST-AUS, including Head ST-AUS (LL-AUS1), Limb ST-AUS (LL-AUS2), Leg ST-AUS (LL-AUS3), and Foot ST-AUS (LL-AUS4).

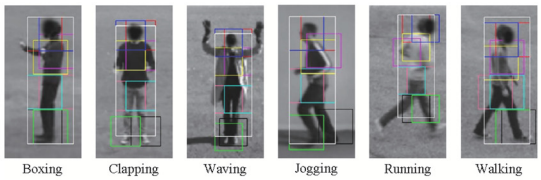
Middle Level (ML) ST-AUS: ML ST-AUS represent the composite partwise visual dynamics. The space-time interest points in head/limb regions can be integrated into the same group to represent the upper-body characteristics while the space-time interest points in leg/foot regions can be integrated into the other group to represent the lower-body characteristics. Consequently, we can get the Upper ST-AUS (ML-AUS1) and Lower ST-AUS (ML-AUS2) (Fig 1f-1g), as we do for LL ST-AUS representation.

High Level (HL) ST-AUS: HL ST-AUS focuses on the global visual dynamics. The space-time interest points of one person are considered as one group for codebook learning. The feature representation of the entire body, Full ST-AUS (HL-AUS) (Fig 1h), can be generated as we do for LL/ML ST-AUS representation.

LL ST-AUS, ML ST-AUS and HL ST-AUS together form the ST-AUS representation, ST-AUS = {LL-AUS1, LL-AUS2, LL-AUS3, LL-AUS4, ML-AUS1, ML-AUS2, HL-AUS}, for one action video, which conveys different partwise temporal structures.

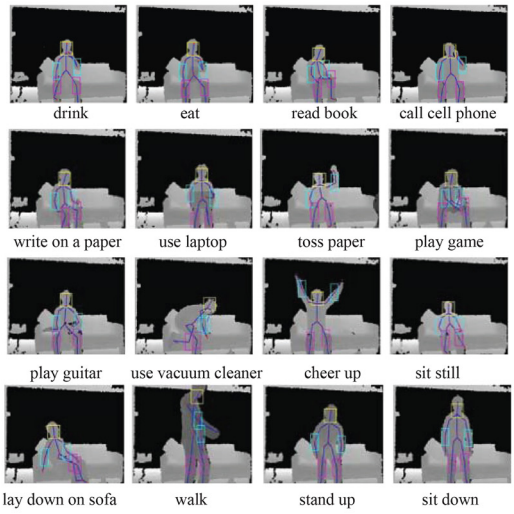
Multi-task Conditional Random Fields

In this section, we respectively present the probabilistic model of the multi-task conditional random fields (MTCRFs) model and the methods for MTCRFs learning and inference.



(a) KTH

(b) TJU



(c) MDA

Fig 2. Samples from KTH (a), TJU (b), and MDA (c).

doi:10.1371/journal.pone.0130884.g002

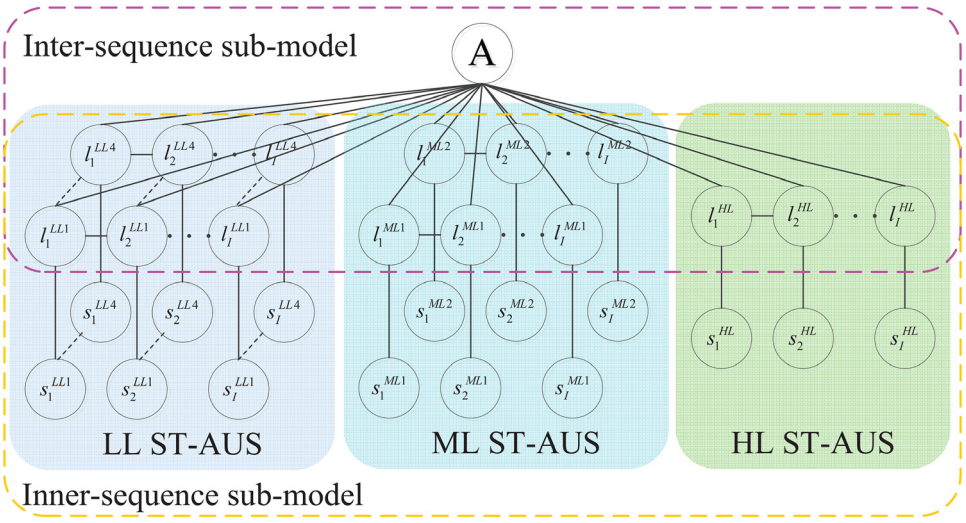


Fig 3. Graph structure of MTCRFs.

doi:10.1371/journal.pone.0130884.g003

Probabilistic Model of MTCRFs

We design the specific graph structure $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ (Fig 3) for the MTCRFs model. $\mathcal{V}$ means the node set, including both observation nodes and hidden state nodes. $\mathcal{E}$ means the edge set, including the transition between adjacent hidden states and the correlation between the hidden state and the action label. In terms of the designed graph structure, each action video, depicted by the extracted ST-AUS representation, can be represented by P parallel sequences $S = \{s^p\}_{p=1}^{P}$ with the chain structure, where $s^p = \{s_i^p\}_{i=1}^{I}$ denotes individual part-induced ST-AUS which represents the temporal dynamics of a specific body part during one action. S is assigned with a specific action label $A \in \mathcal{A}$. To model the state transition within individual ST-AUS (s^p), we utilize the hidden state layer $L = \{l^p\}_{p=1}^{P} (L \in \mathcal{L})$ to correlate the adjacent observations, in which $l^p = \{l_i^p\}_{i=1}^{I}$ means the hidden state sequence corresponding to s^p. Each l_i^p is a member of a finite discrete set $\mathcal{L}^p$ of the p^{th} ST-AUS. All the hidden states are correlated by the edge between individual hidden state and the action label node. Consequently, all ST-AUSs can be correlated and will contribute for the modeling and inference of the action category.

To take advantage of both inner-sequence and inter-sequence context for MTCRFs modeling, the proposed probabilistic model of MTCRFs can be formulated with two parts:

$$P(A|S, \Theta) = P_1(A|S, \phi) + P_2(A|S, \psi) \tag{1}$$

where $\Theta = \{\phi, \psi\}$ denotes the weight coefficients of the model; $P_1(A|S, \phi)$ and $P_2(A|S, \psi)$ denotes the inner-sequence sub-model and the inter-sequence sub-model, respectively.

The inner-sequence sub-model $P_1(A|S, \phi)$ can be formulated as follows:

$$P_1(A|S, \phi) = \sum_L P(A, L|S, \phi) = \frac{1}{Z_1} \sum_L \exp\left(\phi^\top \cdot \mu(A, S, L)\right) \tag{2}$$

where Z_1 is the partition function for normalization. The potential function of the inner-sequence sub-model, $\phi^\top \cdot \mu(A, S, L) \in \mathbb{R}$, can be defined with the first-order attribute between the observation and the corresponding hidden state and the second-order attribute between

adjacent hidden states:

$$\phi^{\top} \cdot \mu(A,S,L) = \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{k=1}^{K_1}\phi_{1,k}F_{1,k}(l_i^p, s_i^p) + \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{k=1}^{K_2}\phi_{2,k}F_{2,k}(l_i^p, l_{i+1}^p) \tag{3}$$

where the first term denotes the correlation between the observation s_i^p and the corresponding hidden state l_i^p and the second term denotes the transition between adjacent hidden states. $F_{1,k}(l_i^p, s_i^p) \in \mathbb{R}$ is a general expression of the first order attribute function which represents the relationship between pairwise observation and hidden state nodes. $F_{2,k}(l_i^p, l_{i+1}^p) \in \mathbb{R}$ is a general expression of the second order attribute function which represents the relationship between pairwise hidden state nodes. Different from previous work [30] which only consider the first order attribute between the observation node and the corresponding hidden state node, we leverage both first order and second order attributes to formulate the inner-sequence sub-model. It is intuitive that different actions might have the similar action unit. For example, both boxing and handclapping can have similar hand/limb motion. It is not discriminative enough to distinguish each other only with the body motion during one state. However, it will be more easily to discriminate both when considering the motion changes between adjacent states since there will be different motion intensity, direction, etc. These motion changes can be well represented by the second-order attribute.

The inter-sequence sub-model $P_2(A|S, \psi)$ can be likewise formulated as follows:

$$P_2(A|S,\psi) = \sum_{L} P(A,L|S,\psi) = \frac{1}{Z_2}\sum_{L}\exp\left(\psi^{\top} \cdot v(A,S,L)\right) \tag{4}$$

where Z_2 is another partition function for normalization. The potential function of the inter-sequence sub-model, $\psi^{\top} \cdot v(A, S, L) \in \mathbb{R}$, can be defined with the first-order attribute between the sequence class and the corresponding hidden state and the second-order attribute between the sequence class and the pairwise hidden states:

$$\psi^{\top} \cdot v(A,S,L) = \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{A' \in \mathcal{A}}\sum_{k=1}^{K_3}\psi_{1,k}F'_{1,k}(l_i^p, A') + \sum_{p=1}^{P}\sum_{i=1}^{I}\sum_{A' \in \mathcal{A}}\sum_{k=1}^{K_4}\psi_{2,k}F'_{2,k}(l_i^p, l_{i+1}^p, A') \tag{5}$$

where the first term denotes the correlation between the sequence label A and individual hidden state l_i^p and the second term denotes the correlation between the sequence label A and pairwise hidden states. $F'_{1,k}(l_i^p, A') \in \mathbb{R}$ is the first order attribute function which represents the relationship between the sequence label and individual hidden state node. $F'_{2,k}(l_i^p, l_{i+1}^p, A') \in \mathbb{R}$ is the second order attribute function which represents the relationship between the sequence label and pairwise hidden state nodes. Since all the hidden states are linked by the sequence label node, this model can learn the latent correlation among different ST-AUSs. This is quite understandable since each action has specific motion pattern and the movement of different body parts during one action is implicitly constrained by each other. Different from previous work [31] which designed the edges between the pairwise hidden states from two different chain-structured sequences to learn the inter-sequence correlation, we omit this edges due to

two reasons: 1) there might exist asynchrony among different ST-AUSs since different part-induced ST-AUSs can have different motion dynamics. It is unreasonable to impose strong constraint by linking the pairwise hidden stated with the same index from two sequences. 2) the inter-sequence edges linking pairwise hidden states will significantly increase the complexity of the graph structure and directly make model learning much more difficult.

MTCRFs Learning and Inference

Since both sub-models have the similar formulation in the exponential family manner, they can be separately optimized by the gradient descent method [32]. Here we take the inner-sequence sub-model, $P_1(A|S, \phi)$, as an example to illustrate the model learning method. Model learning of $P_1(A|S, \phi)$ can be accomplished by maximize the following likelihood objective function, given the training set with N samples $\Gamma = \{(S^i, A^i)\}_{i=1}^N$:

$$\mathcal{L}(\Gamma;\phi) = \sum_{(S_i,A_i)\in\Gamma} log \sum_L \exp\left(\phi^\top \cdot \mu(A,S,L)\right)/Z_1 - \sum_{k=1}^{K} \frac{1}{2\sigma^2}\|\phi_k\|^2 \tag{6}$$

where we suppose $\phi \in \mathbf{R}^K$ ($K = K_1+K_2$). The objective function $\mathcal{L}(\Gamma;\phi)$ is the summation of log-likelihood of all training samples minus a regularization term. The second term $\frac{1}{2\sigma^2}\|\phi\|^2$ is a an L_2-norm regularization when parameters are assumed to be Gaussian distributed with variance σ^2. It is set to avoid overfitting.

To achieve the optimal parameter ϕ^*, we can take a partial derivative of $\mathcal{L}(\Gamma;\phi)$ with respect to each entity ϕ_k as:

$$\frac{\partial \mathcal{L}(\Gamma;\phi)}{\partial \phi_k} = \sum_{(S_i,A_i)\in\Gamma} \frac{\partial log\sum_L \exp\left(\phi^\top \cdot \mu(A,S,L)\right)/Z_1}{\partial \phi_k} - \frac{\phi_k}{\sigma^2} \tag{7}$$

Since $\phi_{1,\,k}$ only exists in the first-order attribute term, the partial derivative of the core of Eq 7 can be derived as:

$$\begin{aligned}
\frac{\partial log\sum_L \exp\left(\phi^\top \cdot \mu(A,S,L)\right)/Z_1}{\partial \phi_{1,k}} = & \frac{\partial log\sum_L \exp\left(\phi^\top \cdot \mu(A,S,L)\right)}{\partial \phi_{1,k}} - \\
& \frac{\partial log\sum_{A'\in\mathcal{A}}\sum_L \exp\left(\phi^\top \cdot \mu(A',S,L)\right)}{\partial \phi_{1,k}} \\
= & \sum_L \sum_{p=1}^{P} \sum_{i=1}^{I} P(L|S_i,A_i;\phi)\cdot F_{1,k}(l_i^p, s_i^p) - \\
& \sum_{A'} \sum_L \sum_{p=1}^{P} \sum_{i=1}^{I} P(L,A'|S_i;\phi)\cdot F_{1,k}(l_i^p, s_i^p) \\
= & \; p_{1,k}(A,L,S;\phi)
\end{aligned} \tag{8}$$

In the same way, $\phi_{2,k}$ only exists in the second attribute term and the partial derivative of the core of Eq 7 is:

$$\begin{aligned}\frac{\partial log\sum_L \exp(\phi^\top \cdot \mu(A,S,L))/Z_1}{\partial \phi_{2,k}} &= \sum_L \sum_{p=1}^{P} \sum_{i=1}^{I} P(L|S_i,A_i;\phi)\cdot F_{2,k}(l_i^p, l_{i+1}^p) - \\ &\quad \sum_{A'} \sum_L \sum_{p=1}^{P} \sum_{i=1}^{I} P(L,A'|S_i;\phi)\cdot F_{2,k}(l_i^p, l_{i+1}^p) \\ &= p_{2,k}(A,L,S;\phi)\end{aligned} \tag{9}$$

With the partial derivative of Eq 7 with respect to both $\phi_{1,k}$ and $\phi_{2,k}$ above, the parameters can be updated as follows:

$$\begin{aligned}\phi_{1,k}^t &= \phi_{1,k}^{t-1} + \gamma \frac{\partial \mathcal{L}(\Gamma,\phi)}{\partial \phi_{1,k}}|_{\phi=\phi^{t-1}} \\ &= \phi_{1,k}^{t-1} + \gamma \left(\sum_{(S_i,A_i)\in\Gamma} p_{1,k}(A,L,S;\phi) - \frac{\phi_{1,k}^{t-1}}{\sigma^2} \right)\end{aligned} \tag{10}$$

$$\begin{aligned}\phi_{2,k}^t &= \phi_{2,k}^{t-1} + \gamma \frac{\partial \mathcal{L}(\Gamma,\phi)}{\partial \phi_{2,k}}|_{\phi=\phi^{t-1}} \\ &= \phi_{2,k}^{t-1} + \gamma \left(\sum_{(S_i,A_i)\in\Gamma} p_{2,k}(A,L,S;\phi) - \frac{\phi_{2,k}^{t-1}}{\sigma^2} \right)\end{aligned} \tag{11}$$

where t means the iteration index; γ means the learning rate.

The parameter ψ of the inter-sequence sub-model $P_2(A|S,\psi)$ can be optimized in the same way and we omit the related derivation. For model inference, the optimal sequence label A^* can be predicted with the belief propagation algorithm [33]:

$$A^* = \arg\max_{A\in\mathcal{A}} P(A|S;\Theta^*) \tag{12}$$

where $\Theta^* = \{\phi^*, \psi^*\}$.

Experiment Method

Data

The proposed method is evaluated on the popular datasets in both RGB and depth modalities, including KTH, TJU, MSR Daily Activity 3D (MDA) respectively shown in Fig 2a, 2b and 2c. We will briefly introduce the datasets as follows.

- **KTH**: KTH [22] contains 6 action categories: 1. boxing, 2. hand clapping, 3. hand waving, 4. jogging, 5. running and 6. walking. Each action was performed by 25 people in 4 different scenarios, including indoor, outdoor, changes in clothing and variations in scale. Each video sample contains one subject engaged in a single activity in a certain condition. For fair comparison, we strictly followed the experiment setting as [34] in the "split" manner.
- **MDA**: MSR Daily Activity 3D [35] is very challenging since it captures the human daily activities with human-object interaction. It consists of 16 action(1.drink, 2.eat, 3.read book, 4.call cell phone, 5.write on a paper, 6.use laptop, 7.toss paper, 8.play game, 9.play guitar, 10.

use vacuum cleaner, 11.cheer up, 12.sit still, 13.lay down on sofa, 14.walk, 15.stand up, 16.sit down). Each person performs one action in two scenarios, sitting on sofa and standing. We eliminated 4 foot joint positions and only used the other 16 key joints for partwise BoW generation because the foot regions are usually out of the range of Kinect. The Leave-One-Subject-Out strategy is leveraged for evaluation.

- **TJU**: We contributed a public dataset, TJU, to provide an action dataset with more samples and action categories. TJU was recorded in the front view and RGB image(640 × 480), depth image(640 × 480) and skeleton were recorded. The dataset contains 22 action categories: 1. walking, 2. jogging, 3. running, 4. boxing, 5. waving, 6. clapping, 7. bend, 8. jacks, 9.jump, 10. p-jump, 11. side, 12. single-wave, 13. draw-x, 14. draw-tick, 15. draw-circle, 16. kick, 17. side-kick, 18. tennis-swing, 19. tennis-serve, 20. side-box, 21. p-bend and 22. sit-down. Each of the 22 actions was performed four times by 20 people. There are totally 1760 samples. The dataset is splitted into two parts, the first 12 subjects for training and the rest for test. TJU is challenging since it contains the more complex action samples with high intra variation and was captured in both light and dark environments.

Implementation Details

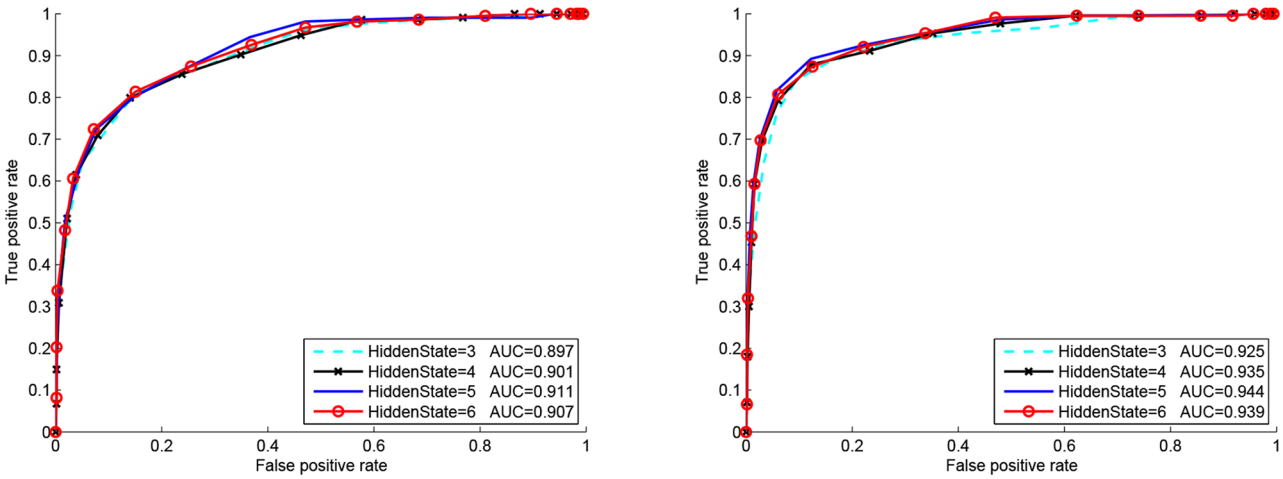
For RGB datasets (KTH and TJU), we utilized STIP [36] for spatiotemporal interest point localization and representation because it has shown superior performances in the BoW+SVM framework [34]. For depth dataset (MDA), the local HON4D descriptor and the 3Djoint position feature were utilized. As Oreifej [37], we concatenated HON4D and 3Djoint position features for representation and further utilized them for dictionary learning and BoW representation. The original parameter settings were implemented to extract STIP and HON4D features for fair comparison. For KTH, the part-based model [17] was implemented on each frame to localize body part regions. For TJU and MDA, we directly used the skeleton data provided in each dataset for body part localization. Then the space-time feature points were grouped into different body parts in terms of the geometric distances between the point and the center of each part region. To construct the visual vocabulary of individual part, we clustered 20,000 feature points of each part region sampled from the training videos with the K-means algorithm. The number of visual words was empirically set with 100, which showed satisfactory performance for partwise ST-AUS representation. The ST-AUS in our experiment lasted for 30 frames in temporal domain and the overlap between two adjacent ST-AUSs was set with 15.

Experimental Results

To show the superiority of the proposed method, the following two comparison experiments were implemented.

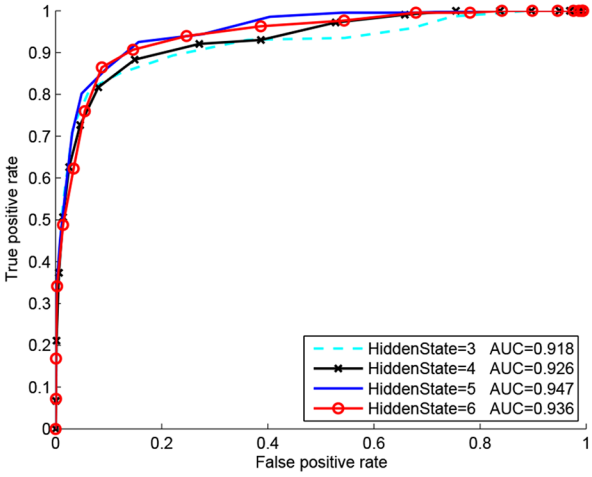
Comparison against other fusion methods

With the ST-AUS representation, the proposed MTCRFs model can be trained and utilized for action recognition based on the Maximum A Posteriori criteria. To select the best hidden state for temporal modeling, we plotted the ROC curve of each hidden state number and the best parameter can be selected when the area under curve (AUC) of the corresponding ROC reached the maximum. In our experiments, we varied the number of hidden states from 3 to 6 per chain for parameter selection. From Figs 4, 5 and 6, it is obvious that the MTCRFs model on each dataset can achieve the best performance with *hidden_state* = 5.

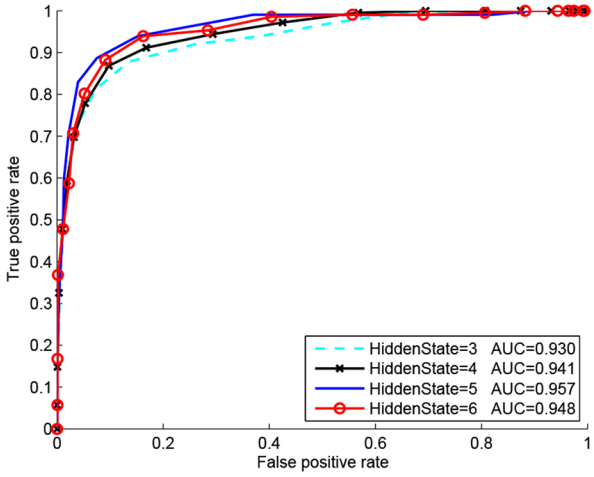


(d) HL-HCRF

(e) ML-mcDRFs



(f) LL-mcDRFs



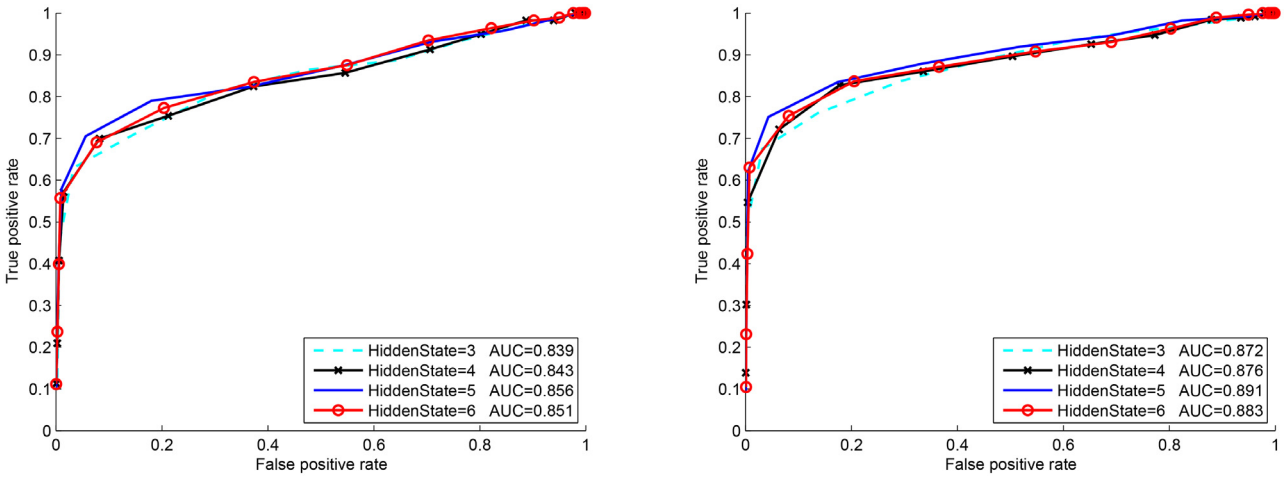
(g) All-mcDRFs

Fig 4. ROC and AUC of KTH with different hidden states.

doi:10.1371/journal.pone.0130884.g004

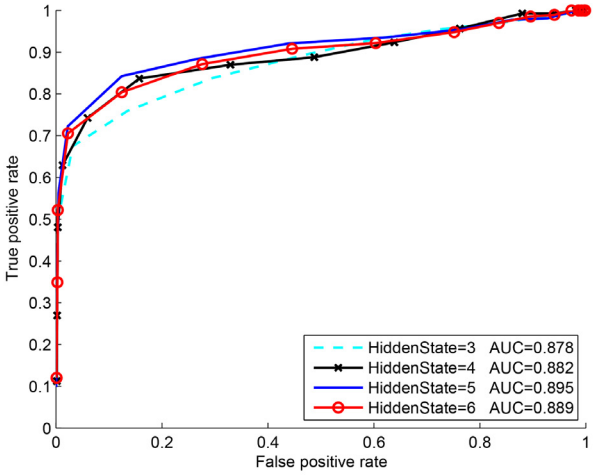
To show the proposed method can benefit sequential dynamic learning, we implemented the following strategies for comparison:

- HL-HCRF: HCRF model trained only with HL-AUS (as shown in Fig 1(h)), which utilized the bodywise feature for temporal modeling.
- LL-DDF: Direct decision-level fusion by linearly combining the posterior probabilities of HCRF models, trained with LL-AUS1, LL-AUS2, LL-AUS3, LL-AUS4 respectively, with equal weights.
- ML-DDF: Direct decision-level fusion by linearly combining the posterior probabilities of HCRF models, trained with ML-AUS1 and ML-AUS2 respectively, with equal weights.

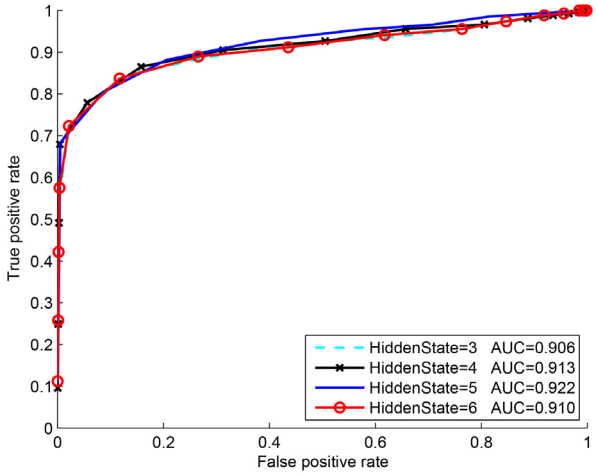


(h) HL-HCRF

(i) ML-mcDRFs



(j) LL-mcDRFs

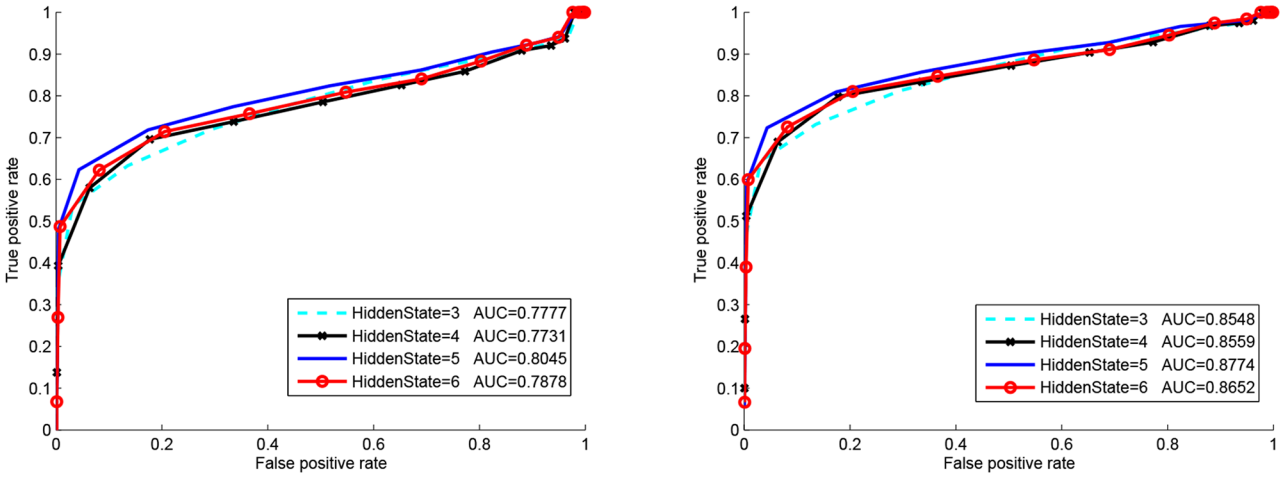


(k) All-mcDRFs

Fig 5. ROC and AUC of TJU with different hidden states.

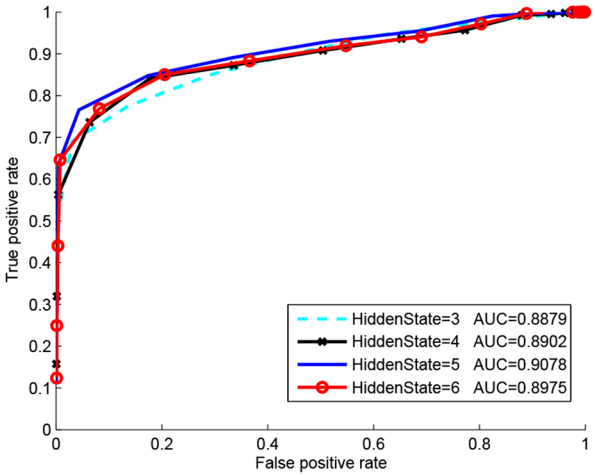
doi:10.1371/journal.pone.0130884.g005

- All-DDF: Direct decision-level fusion by linearly combining the posterior probabilities of HCRF models, trained with 7 components of ST-AUS respectively, with equal weights.
- LL-DFF: HCRF model trained with direct feature-level fusion of LL-AUS1, LL-AUS2, LL-AUS3, LL-AUS4.
- ML-DFF: HCRF model trained with direct feature-level fusion of ML-AUS1 and ML-AUS2.
- All-DFF: HCRF model trained with direct feature-level fusion of 7 components of ST-AUS.
- LL-MTCRFs: MTCRFs model trained with LL-AUS1, LL-AUS2, LL-AUS3, LL-AUS4 as sequential observations.
- ML-MTCRFs: MTCRFs model trained with ML-AUS1 and ML-AUS2 as sequential observations.

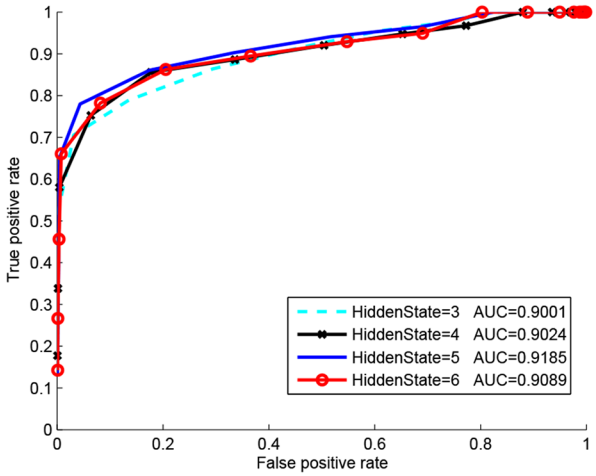


(l) HL-HCRF

(m) ML-mcDRFs



(n) LL-mcDRFs



(o) All-mcDRFs

Fig 6. ROC and AUC of MDA with different hidden states.

doi:10.1371/journal.pone.0130884.g006

- All-MTCRFs: MTCRFs model trained with 7 components of ST-AUS as sequential observations.

The comparison results are shown in Table 1. The confusion matrixes of the optimal performance on three datasets by All-MTCRFs are respectively shown in Fig 7a, 7b and 7c.

From Table 1 we can see that HL-HCRF worked in a traditional manner for HCRF learning simply with global feature and only acieved 90.0%, 80.5% and 71.2% for KTH, TJU, and MDA respectively. ML-DDF, LL-DDF, and All-DDF worked in a decision-fusion manner and cannot take advantage of the temporal context of all sequences for sequential structure learning. It can only improve the performance slightly. Comparatively, ML-DFF, LL-DFF, and All-DFF worked in a feature-fusion manner and can leverage the temporal context of all partwise

Table 1. Performance Comparison on KTH, TJU, and MDA (%).

Method	KTH	TJU	MDA
HL-HCRF	90.0	80.5	71.2
ML-DDF	91.0	81.4	75.7
LL-DDF	91.4	82.6	77.4
All-DDF	91.9	84.3	78.1
ML-DFF	91.4	84.7	81.6
LL-DFF	91.9	85.2	81.9
All-DFF	92.4	86.8	84.4
ML-MTCRFs	92.9	86.2	86.7
LL-MTCRFs	93.3	88.1	88.0
All-MTCRFs	**94.8**	**90.5**	**89.8**

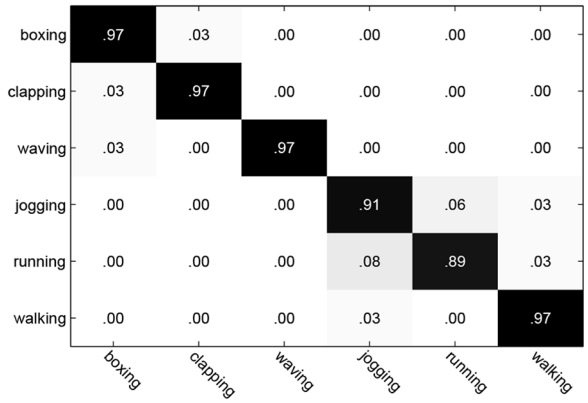
doi:10.1371/journal.pone.0130884.t001

sequences for sequential structure learning. Consequently the feature-level fusion method can further augment the performance. Since the proposed MTCRFs can preserve and learn the temporal context of individual ST-AUS while relaxing the interaction inbetween to handle the asynchronization, it can achieve the best accuracy of 94.8%, 90.5% and 89.8% on three datasets respectively and consistently outperform all the others.

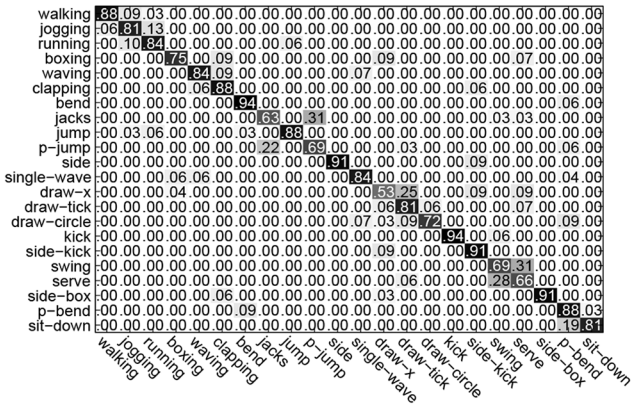
The action category-wise comparison among HL-HCRF, LL-MTCRFs, ML-MTCRFs, and All-MTCRFs for three datasets are listed in Fig 8a, 8b and 8c. For KTH, All-MTCRFs can work better or equally to the others in 4 out of 6 actions. Especially it can drastically improve the accuracy of Jogging from around 80% to 91.2%, which is the most challenging one in KTH. For TJU, All-MTCRFs can rank 1st in 17 out of 22 action. It can augment the performance of 4 actions (clapping, jacks, p-jump, draw-circle) with more than 10% accuracy. For MDA in the depth modality, All-MTCRFs can also improve the performance and rank 1st in 14 out of 16 action. It can augment the performance of 6 actions (drink, read book, write on a paper, sit still, toss paper, lie down on sofa) with more than 20% accuracy. This comparison demonstrates that the proposed method is more robust to high intra variation caused by more complex actions and diverse environments. The action-wise comparison among the first three methods in Fig 8 shows that LL-MTCRFs can work better or equally to HL-HCRF and ML-MTCRFs in 3 out of 6 actions in KTH, 16 out of 22 actions in TJU, 12 out of 16 actions in MDA. It further demonstrates that the body part-induced ST-AUS representation is more discriminative for local dynamic description and consequently facilitates human action recognition.

Comparison against state-of-the-art methods

To show the superiority of the proposed method, we compared its performances on KTH, TJU, and MDA to the state-of-the-art methods. Tables 2 and 3 show the comparison on KTH & MDA. The performances of the competing methods are cited from the corresponding papers. From Tables 2 and 3 it is obvious that the proposed method can outperform the competing methods in both RGB and depth modalities. Especially, the proposed method can outperform the slow feature-based method [47], which can also explore the temporal context. Theoretically, Zhang *et al.* [47] utilized the slow feature learning for frame-wise feature transformation and further computed the accumulated squared derivative (ASD) feature of video-wise representation by leveraging the statistical characteristics of frame-wise slow features. Although the ASD feature also conveys the temporal characteristics, it loses the multiple sequential context



(a) KTH



(b) TJU

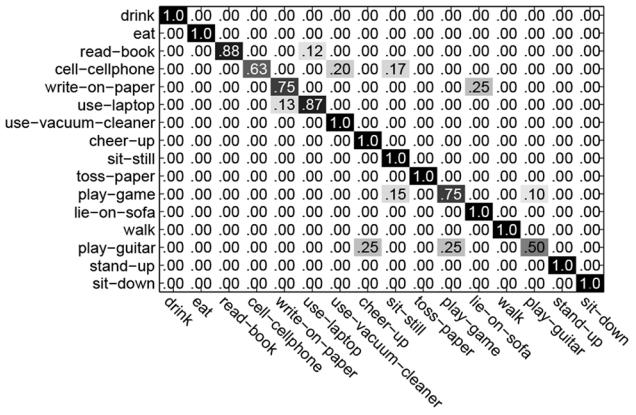
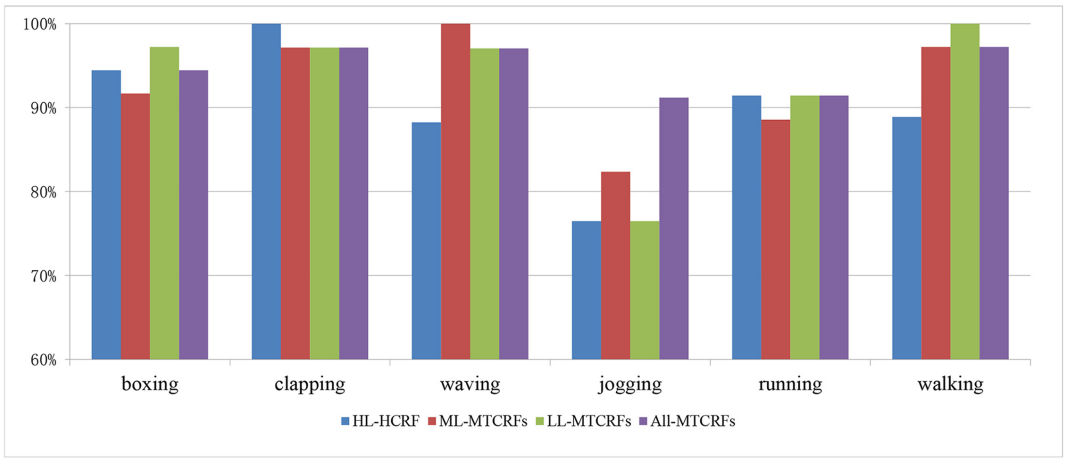
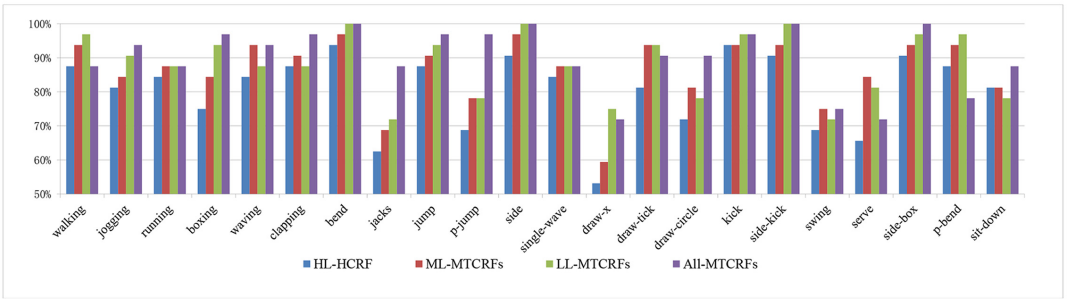


Fig 7. The confusion matrixes of All-MTCRFs on KTH, TJU, and MDA.

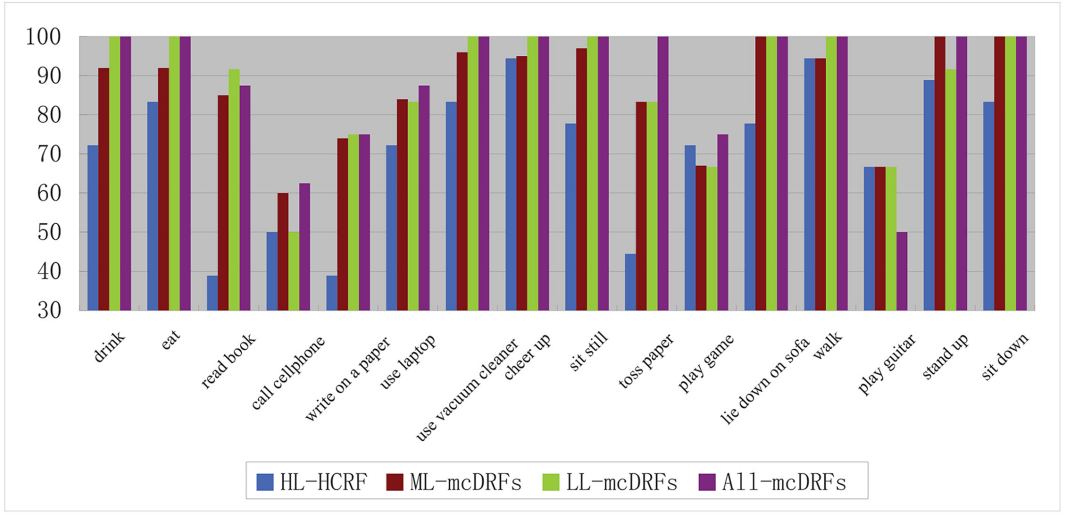
doi:10.1371/journal.pone.0130884.g007



(s) KTH



(t) TJU



(u) MDA

Fig 8. Action category-wise comparison among HL-HCRF, LL-MTCRFs, ML-MTCRFs, and All-MTCRFs on KTH, TJU, and MDA.

doi:10.1371/journal.pone.0130884.g008

Table 2. Comparison with state of the arts on KTH.

Method	Year	Accuracy(%)
Fathi *et al*.	(CVPR2008) [38]	90.5
Niebles *et al*.	(IJCV2008) [39]	83.3
Laptev *et al*.	(CVPR2008) [28]	91.8
Kläser *et al*.	(BMVC2008) [3]	91.4
Wang *et al*.	(BMVC2009) [34]	92.1
Gilbert *et al*.	(ICCV2009) [40]	94.5
Taylor *et al*.	(ECCV2010) [41]	90.0
Kovashka *et al*.	(CVPR2010) [42]	94.5
Le *et al*.	(CVPR2011) [43]	93.9
Wang *et al*.	(CVPR2011) [44]	93.8
Minhas *et al*.	(TCSVT2012) [45]	94.4
Ballan *et al*.	(TMM2012) [46]	92.7
U-SFA	(TPAMI2012) [47]	84.7
S-SFA	(TPAMI2012) [47]	88.8
D-SFA	(TPAMI2012) [47]	91.2
SD-SFA	(TPAMI2012) [47]	93.5
Ji *et al*.	(TPAMI2013) [48]	90.2
Ma *et al*.	(TCSVT2013) [49]	94.4
Proposed		**94.8**

doi:10.1371/journal.pone.0130884.t002

Table 3. Comparison with state of the arts on MDA.

Method	Year	Accuracy(%)
HOG	CVPR2008 [28]	79.1
LOP	CVPR2012 [50]	42.5
Joint Position	CVPR2012 [50]	68.0
Actionlet Ensemble	CVPR2012 [50]	85.8
DCSF	CVPR2013 [51]	83.6
DCFS+Joint	CVPR2013 [51]	88.2
Proposed		**90.5**

doi:10.1371/journal.pone.0130884.t003

represented by the proposed part-induced spatiotemporal action unit sequence. Therefore, it is reasonable that the proposed method can outperform the slow feature-based method.

Table 4 shows the comparison on TJU. Since TJU is a new dataset prepared by our group and no off-the-shelf comparison available, we re-implemented six representative methods for comparison: 1) STIP-BoW+SVM: the standard BoW+SVM method with the STIP feature [28]; 2) 3DHOG-BoW+SVM: the standard BoW+SVM method with the 3DHOG feature [3]; 3) BoW+SRC: the standard BoW representation with STIP and sparse representation-based classification [52]; 4) BoW+CRC: the standard BoW representation with STIP and collaborative representation-based classification [53]; 5) HL-AUS+HMM: bodywise action unit sequence with STIP and hidden markov model [54]; 6) HL-AUS+semi-DRF: bodywise action unit sequence with STIP and semi-Markov discriminative random fields [55]. The first four methods belong to the popular BoW-based methods for classification without utilizing temporal

Table 4. Comparison with state of the arts on TJU.

Method	Accuracy(%)
STIP-BoW+SVM	83.3
3DHOG-BoW+SVM	82.6
BoW+SRC	83.9
BoW+CRC	84.6
HL-AUS+HMM	85.4
HL-AUS+semi-DRF	87.0
Proposed	**90.5**

doi:10.1371/journal.pone.0130884.t004

context. Therefore it is expected that their performances are relatively lower than the other temporal inference methods. HMM and semi-DRF are two popular models which leverage sequence structure for temporal modeling. HMM is theoretically limited by its inability to accommodate long-range dependencies among observations or multiple overlapping features because they assume that the observations are conditionally independent. Therefore semi-DRF works better than HMM. However, both HMM and semi-DRF cannot take advantage of multiple partwise ST-AUS simultaneously for sequence structuring learning and the latent correlation discovery. Consequently, the proposed MTCRFs model can achieve the best performance of 90.5% accuracy.

Conclusion

In this paper we propose a novel human action recognition method based on multi-task conditional random fields model. For feature representation, we propose the part-induced spatio-temporal action unit sequence to represent each action sample with multiple partwise sequential feature subspaces. For model learning, we propose the multi-task conditional random fields (MTCRFs) model to discover the sequence-specific structure and the sequence-shared relationship. Specifically, we propose the probabilistic model and the corresponding learning and inference methods to discover temporal context within individual action unit sequence and the latent correlation among different body parts. The comparison experiments demonstrated its superiority on two RGB human action datasets, KTH & TJU, and one depth dataset, MSR Daily Activity 3D.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (61472275, 61170239), the Tianjin Research Program of Application Foundation and Advanced Technology (15JCYBJC16200), the grant of Elite Scholar Program of Tianjin University (2014XRG-0046).

Author Contributions

Conceived and designed the experiments: AAL YTS WZN ZXY. Performed the experiments: AAL WZN. Analyzed the data: AAL YTS WZN ZXY. Wrote the paper: AAL YTS WZN ZXY.

References

1. Aggarwal JK, Ryoo MS (2011) Human activity analysis: A review. ACM Comput Surv 43: 1–43. doi: 10.1145/1922649.1922653

2. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008 IEEE Conference on. IEEE, pp. 1–8.
3. Klaser A, Marszalek M, Schmid C (2008) A spatio-temporal descriptor based on 3 D-gradients. In: BMVC'08.
4. Weinland D, Boyer E, Ronfard R (2007) Action recognition from arbitrary views using 3d exemplars. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14–20, 2007. pp. 1–7.
5. Ivanov YA, Bobick AF (2000) Recognition of visual activities and interactions by stochastic parsing. IEEE Trans Pattern Anal Mach Intell 22: 852–872. doi: 10.1109/34.868686
6. Ryoo MS, Aggarwal JK (2006) Recognition of composite human activities through context-free grammar based representation. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17–22 June 2006, New York, NY, USA. pp. 1709–1718.
7. Pei M, Jia Y, Zhu S (2011) Parsing video events with goal inference and intent prediction. In: IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6–13, 2011. pp. 487–494.
8. Hamid R, Maddi S, Johnson AY, Bobick AF, Essa IA (2009) A novel sequence representation for unsupervised analysis of human activities. Artif Intell 173: 1221–1244. doi: 10.1016/j.artint.2009.05.002
9. Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML'01.
10. Lv F, Nevatia R (2006) Recognition and segmentation of 3-d human action using hmm and multi-calss adaboost. In: Proc. ECCV. volume 3954, pp. 359–372.
11. Han L, Wu X, Liang W, Hou G, Jia Y (2010) Discriminative human action recognition in the learned hierarchical manifold space. Image and Vision Computing 28: 836–849. doi: 10.1016/j.imavis.2009.08.003
12. Quattoni A, Wang S, Morency L, Collins M, Darrell T (2007) Hidden conditional random fields. Pattern Analysis and Machine Intelligence 29: 1848–1852. doi: 10.1109/TPAMI.2007.1124
13. Morency L, Quattoni A, Darrell T (2007) Latent-dynamic discriminative models for continuous gesture recognition. In: CVPR'07. IEEE, pp. 1–8.
14. Liu A (2012) Bidirectional integrated random fields for human behavior understanding. Electronics Letters 48: 262–U1556. doi: 10.1049/el.2011.3530
15. Liu A, Li K, Kanade T (2012) A semi-markov model for mitosis segmentation in time-lapse phase contrast microscopy image sequences of stem cell populations. IEEE Trans Med Imaging: 359–369.
16. Liu A, Xu N, Su Y, Lin H, Hao T (2015) Single/multi-view human action recognition via regularized multi-task learning. Neurocomputing 151: 544–553. doi: 10.1016/j.neucom.2014.04.090
17. PFFelzenszwalb, RBGirshick, DMcAllester, DRamanan (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell.
18. YWang, GMori (2011) Hidden part models for human action recognition: Probabilistic versus max margin. IEEE Trans Pattern Anal Mach Intell.
19. Wang Y, Tran D, Liao Z, Forsyth F (2012) Discriminative hierarchical part-based models for human parsing and action recognition. Journal of Machine Learning Research.
20. YTian, RSukthankar, MShah (2013) Spatiotemporal deformable part models for action detection. In: CVPR'13.
21. Liu A, Su Y, Jia P, Gao Z, Hao T, et al. (2015) Multipe/single-view human action recognition via part-induced multi-task structural learning. IEEE Transactions on Cybernetics 45: 1194–1208. doi: 10.1109/TCYB.2014.2347057 PMID: 25167566
22. CSchuldt, ILaptev, BCaputo (2004) Recognizing human actions: a local SVM approach. In: ICPR (3)'04. IEEE, volume 3, pp. 32–36.
23. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: CVPR'12. pp. 1290–1297.
24. Chen X, Pan W, Kwok JT, Carbonell JG (2009) Accelerated gradient method for multi-task sparse learning problem. In: ICDM'09. pp. 746–751.
25. Willems G, Tuytelaars T, Gool LJV (2008) An efficient dense and scale-invariant spatio-temporal interest point detector. In: ECCV'08. pp. 650–663.
26. Scovanner P, Ali S, Shah M (2007) A 3-dimensional sift descriptor and its application to action recognition. In: ACM Multimedia'07.
27. Jhuang H, Serre T, Wolf L, Poggio T (2007) A biologically inspired system for action recognition. In: ICCV'07. pp. 1–8.

28. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: CVPR'08.
29. Kläser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3 d-gradients. In: British Machine Vision Conference. pp. 995–1004.
30. Liu A, Nie W, Su Y, Ma L, Hao T (2015) Coupled hidden conditional random fields for RGB-D human action recognition. Signal Processing 112: 74–82. doi: 10.1016/j.sigpro.2014.08.038
31. Su Y, Ma L, Liu A, Yang Z (2014) Max margin discriminative rndom fields for multimodal human action recognition. Electronics Letters 50: 870–872. doi: 10.1049/el.2014.1027
32. Quattoni A, Wang S, Morency LP, Collins M, Darrell T (2007) Hidden conditional random fields 29: 1848–1852.
33. Welling M (2004) On the choice of regions for generalized belief propagation. In: UAI'04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, Banff, Canada, July 7–11, 2004. pp. 585–592.
34. Wang H, Ullah MM, Klaser A, Laptev I, Schmid C (2009) Evaluation of local spatio-temporal features for action recognition. In: BMVC'09.
35. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: CVPR'12. pp. 1290–1297.
36. Laptev I, Lindeberg T (2003) Space-time interest points. In: ICCV'03. pp. 432–439.
37. Oreifej O, Liu Z (2013) Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: CVPR'13.
38. Fathi A, Mori G (2008) Action recognition by learning mid-level motion features. In: CVPR'08.
39. Niebles JC, Wang H, Li FF (2008) Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision: 299–318. doi: 10.1007/s11263-007-0122-4
40. Gilbert A, Illingworth J, Bowden R (2009) Fast realistic multi-action recognition using mined dense spatio-temporal features. In: ICCV'09.
41. Taylor GW, Fergus R, LeCun Y, Bregler C (2010) Convolutional learning of spatio-temporal features. In: ECCV'10. pp. 140–153.
42. Kovashka A, Grauman K (2010) Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: CVPR'10. pp. 2046–2053.
43. Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR'11.
44. Wang J, Chen Z, Wu Y (2011) Action recognition with multiscale spatio-temporal contexts. In: CVPR'11. pp. 3185–3192.
45. Minhas R, Mohammed A, Wu Q (2012) Incremental learning in human action recognition based on snippets. IEEE Trans Circuits Syst Video Techn: 1529–1541.
46. Ballan L, Bertini M, Bimbo A, Seidenari L, Serra G (2012) Effective codebooks for human action representation and classification in unconstrained videos. IEEE Transactions on Multimedia.
47. Zhang Z, Tao D (2012) Slow feature analysis for human action recognition. IEEE Trans Pattern Anal Mach Intell: 436–450.
48. Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell: 221–231.
49. Ma A, Yuen P, Zou W, Lai J (2013) Supervised spatio-temporal neighborhood topology learning for action recognition. IEEE Trans Circuits Syst Video Techn: 1447–1460.
50. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. In: CVPR'12. pp. 1290–1297.
51. Xia L, Aggarwal J (2013) Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: CVPR'13.
52. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell: 210–227.
53. Zhang L, Yang M, Feng X (2011) Sparse representation or collaborative representation: Which helps face recognition? In: ICCV'11.
54. Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. In: Proceedings of the IEEE. pp. 257–286.
55. Liu A (2011) Human action recognition with structured discriminative random fields. Electronics Letters 47: 651–653. doi: 10.1049/el.2011.0880