Learning Spatio-Temporal Features for Action Recognition with Modified Hidden Conditional Random Field

Wanru Xu[1(✉)], Zhenjiang Miao[1], Jian Zhang[2], and Yi Tian[1]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing, China
11112063@bjtu.edu.cn
[2] School of Software, Advanced Analytics Institute, University of Technology, Sydney, Australia

Abstract. Previous work on human action analysis mainly focuses on designing hand-crafted local features and combining their context information. In this paper, we propose using supervised feature learning as a way to learn spatio-temporal features. More specifically, a modified hidden conditional random field is applied to learn two high-level features conditioned on a certain action label. Among them, the individual features can describe the appearance of local parts and the interaction features can capture their spatial constraints. In order to make the best of what have been learned, a new categorization model is proposed for action matching. It is inspired by the Deformable Part Model and the intuition is that actions can be modeled by local features in a changeable spatial and temporal dependency. Experimental result shows that our algorithm can successfully recognize human actions with high accuracies both on the simple atomic action database (KTH and Weizmann) and complex interaction activity database (CASIA).

Keywords: Human action recognition · Spatio-temporal features · HCRF · Changeable spatial-temporal constraint model (CSTCM) · Feature learning

1 Introduction

For human action recognition, there are two important issues: feature extraction and efficient classification. The former is to extract discriminative and robust features to describe actions. The latter is to choose the best category model that uses such type features for corresponding classification. In this paper, we solve the two problems with a unified framework.

In recent years, some "conceptually weaker" models such as applying spatio-temporal descriptors [1] to "bag-of-features" (BOF) [2] have achieved promising results on object detection and even human action recognition. However, one obvious limitation is that interactional information of local features is neglected in such model. An interest point is not isolated but surrounded by its spatio-temporal context. Nowadays, several methods aim to integrate some hand-designed contextual features into it. Bregonzio [3] treats the interest points in a sub-volume as point "cloud".

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part I, LNCS 8925, pp. 786–801, 2015.
DOI: 10.1007/978-3-319-16178-5_55

A histogram in [4] which is called "featuretype×featuretype×relationship" is proposed to capture both appearance and relationship between pairwise visual words. A hierarchical model [5] including three context levels: point-level, intra-trajectory and inter-trajectory, is present to structure the spatio-temporal constraint of interest points. A video is modeled using a collection of both global and segment-level features in [20]. A weakness of such approaches is they must apply specific features, so that it is difficult to extend these features to other datasets. Another disadvantage is that after extracting the designed features, no suitable algorithm is proposed for corresponding classification. The common way is just to concatenate these multiple features to generate a new feature vector for action matching. Such simple concatenation may make them submerge each other. Therefore, in this paper we not only replace the hand-crafted features by learned features with a supervised feature learning algorithm, but also propose a new changeable spatial-temporal constraint categorization model (CSTCM) to make the utmost of what have been learned. We also provide evidence that our feature learning method generalizes to different domains and achieves promising results both on atomic action dataset and multi-persons activity dataset.

Hidden conditional random field (HCRF) is allowed to relax the assumption of conditional independence of the observed data, so any flexible spatial constraints among local features can be modeled. Due to its strong capability of description to spatial interaction, HCRF is widely used to build the structure of local patches in object detection. For action recognition, Wang Y et al. [6, 7] use HCRF to model each frame in the video sequence independently and then obtain the class label for the whole video by majority voting of the labels of its frames. Because action is a natural and continuous process, it is certainly not desirable to describe the action with only one frame. Instead of making a decision for the whole video, a modified HCRF is just applied to learn two compact and semantic representations called individual features and interaction features.

Deformable Part Model (DPM) [8] is now treated as the best method for object detection, which can achieve a two-fold improvement in average to the state-of-the-art. The intuition that objects can be modeled by parts in a deformable configuration provides an elegant framework for representing object categories. Inspired by it, a new categorization model (CSTCM) is proposed which can make use of the two high-level features without submerging each of them. Similar to objects, actions are modeled by local features in a changeable spatial-temporal constraint in this paper. Therefore, our classification model measures not only how similar the appearances of local parts are, but also how much their dependencies change.

The overview of our approach is illustrated in Fig.1. For the atomic action, space-time interest point [1] with HOG\HOF [2] is as the original descriptor. For the interactional activity, five trajectory based motion features [9] are extracted. In the training stage, two high-level features are learned by a modified HCRF. In the testing stage, the individual features are used to measure the similarity of appearance, while the interaction features are applied to estimate the change of constraint through the CSTCM. Both the changes in spatial and temporal domain are calculated for the final action matching.

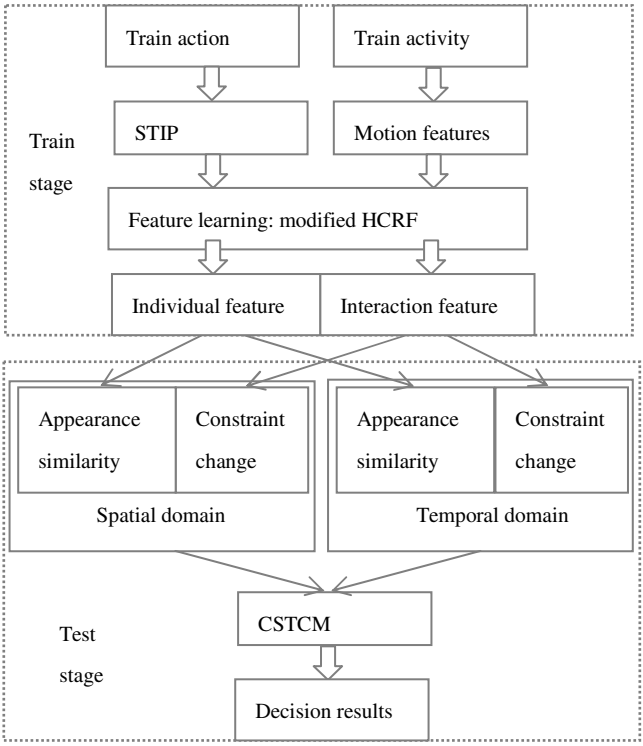


Fig. 1. Overview of our approach

2 Feature Learning

Hidden conditional random field was first proposed for speech classification and then has been applied to gesture [11] and object recognition [10]. In this section, a feature learning algorithm is introduced. By the means of a modified-HCRF, we convert the hand-designed features to data-adaptive descriptors which are more structured and semantic. But most of all, the learned features have the fixed dimension which is convenient for action matching.

Now we describe how to model the *t-th* video frame. **x** is the original feature vector of the *t-th* frame and *y* be the corresponding frame label which is a member of a finite possible action label set *Y*. Our task is not to get a mapping from **x** to *y*, but to learn a series of high-level representations conditioned on the class label *y*. For atomic action recognition, the feature vector **x** composes of a set of local observations $\{x_1, x_2, \ldots, x_N\}$. When the task is to analysis interaction activity, it contains motion features of *N* persons. For any frame **x**, we also define a vector of hidden states $\mathbf{h} = \{h_1, h_2, \ldots, h_N\}$ corresponding to the *N* local patches, where each h_i takes values from a finite possible state set *H*. These variables are not observed, and each of them will assign a state label to the corresponding x_i in training stage. The hidden variables can capture certain un-

derlying spatial structure of these local parts, so they are used as the basic elements to generate two sets of high-level feature descriptors $\{O_1^1, O_2^1, \ldots, O_n^1\}$ and $\{O_1^2, O_2^2, \ldots, O_m^2\}$.

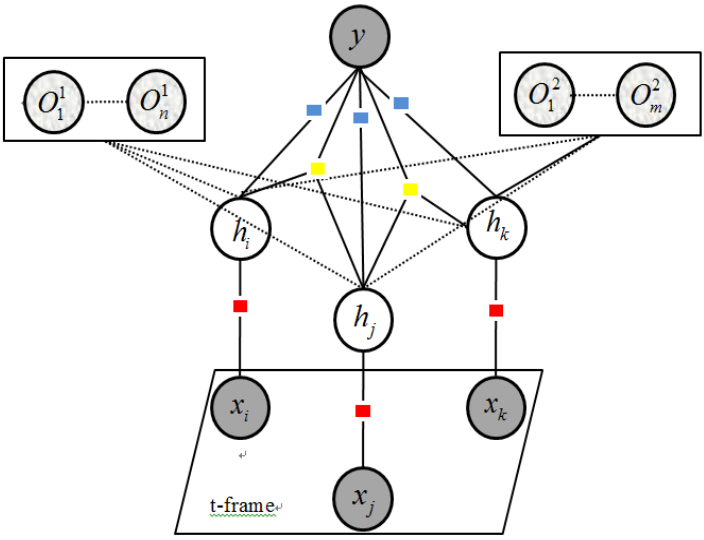


Fig. 2. Illustration of the modified-HCRF

The new modified-HCRF is illustrated in Fig.2. Each circle denotes a variable and each square represents a potential factor. The circles with white are hidden variables, the blacks mean they can be observed and the grays are the high-level features which can be learned from the model. In the graph $G = (E, V)$, $\{h_1, h_2, ..., h_N\}$ are considered to be vertices and edge $(i, j) \in E$ represents the constraint between h_i and h_j. Given $\{\mathbf{x}, y\}$ of *t-th* frame, the modified- HCRF is defined as:

$$p(\mathrm{y} \mid \mathbf{x}; \theta) = \sum_{h \in H} p(\mathrm{y}, \mathrm{h} \mid \mathbf{x}; \theta) = \frac{\sum_{h \in H} e^{\Phi(\mathrm{y},\mathrm{h},\mathbf{x};\theta)}}{\sum_{y \in Y, h \in H} e^{\Phi(\mathrm{y},\mathrm{h},\mathbf{x};\theta)}} \tag{1}$$

Where θ is a set of model parameters and $\Phi(\mathrm{y},\mathrm{h},\mathbf{x};\theta)$ refers to potential function which can measure the compatibility among a class label, a set of observations and a configuration of hidden variables.

2.1 Potential Function

In this paper, the potential function is defined in Eq. (2) which is similar to [6, 7]. Note that the potential is linear in the model parameters $\theta = \{\theta(\mathrm{x}_i, h_i), \theta(y, h_i), \theta(y, h_i, h_j)\}$. $f(x_i)$ refers to original feature of the node i; $f(h_i)$ is the feature corresponding to the

hidden node i; $f(h_i,h_j)$ denotes the feature depending on the edge between node i and j. The details are described in the following.

$$\Phi(\mathrm{y},\mathrm{h},\mathrm{x};\theta)=\sum_{i\in V} f(\mathrm{x}_i)\cdot\theta(\mathrm{x}_i,h_i)+\sum_{i\in V} f(h_i)\cdot\theta(y,h_i)+\sum_{i,j\in E} f(h_i,h_j)\cdot\theta(y,h_i,h_j) \tag{2}$$

Feature-Hidden Potential $f(\mathrm{x}_i)\cdot\theta(\mathrm{x}_i,h_i)$

We use $\theta(\mathrm{x}_i,h_i)$ to refer to the parameter that measures the compatibility between a hidden state h_i and an observational feature $f(x_i)$. The feature-hidden potential describes how likely the local patch x_i is assigned as state h_i. It is formulated as:

$$f(\mathrm{x}_i)\cdot\theta(\mathrm{x}_i,h_i)=\sum_{\mu\in H} f(\mathrm{x}_i)\cdot 1_{\{\mathrm{h}_i=\mu\}}\cdot\theta(\mathrm{x}_i,h_i) \tag{3}$$

Different types of original features are extracted to represent atomic action and interaction activity separately. For simple atomic action, $f(x_i)$ denotes the feature vector describing the appearance of the local patch x_i. Like in [2], we use the Harris 3D detector to detect the local interest points and employ histograms of oriented gradient (HOG) and histograms of optical flow (HOF) to describe the local appearance. Some trajectory based features are applied for complex activity, and now $f(x_i)$ denotes the feature vector describing the motion of the person x_i . Two features in [9] are used here. The moving speed is calculated by Eq. (4) and the motion intensity can be got through Eq. (5).

$$v=\frac{\|\mathrm{L}_1(\mathrm{x},\mathrm{y})-\mathrm{L}_2(\mathrm{x},\mathrm{y})\|}{|t_2-t_1|} \tag{4}$$

$$I=|v_{opticalflow}-v| \tag{5}$$

Where $L_1(x, y)$ and $L_2(x, y)$ refer to the locations of one person at time t_1 and t_2. We use the optical flow to represent the whole motion in the rectangle box of a person, which is defined as the magnitude of the average optical flow speed. It includes two parts, global moving and local motion. The motion intensity I in Eq. (5) is defined as only induced by strong local motions, so we need minus the global moving from the whole motion.

Hidden-label potential $f(h_i)\cdot\theta(y,h_i)$**:**

We use $\theta(y,h_i)$ to denote the parameter that measures the compatibility between a hidden state h_i and an action label y. The hidden-label potential represents how likely the t-th frame contains a local patch with state h_i conditioned on action label y. It can be formulated as:

$$f(h_i)\cdot\theta(y,h_i)=\sum_{\mu\in H,\upsilon\in Y} 1_{\{\mathrm{h}_i=\mu\}}\cdot 1_{\{y=\upsilon\}}\cdot\theta(y,h_i) \tag{6}$$

Edge-Label Potential $f(h_i, h_j) \cdot \theta(y, h_i, h_j)$

As the same, $\theta(y, h_i, h_j)$ corresponds to the parameter for modeling the compatibility between action label *y* and the edge between nodes *i* and *j*. The edge-label potential represents how likely the *t-th* frame contains a pair of local patches with states h_i and h_j conditioned on action label *y*. It is formulated as:

$$f(h_i,h_j)\cdot\theta(y,h_i,h_j)=\sum_{\mu\in H,\omega\in H,\upsilon\in Y} f(h_i,h_j)\cdot 1_{\{h_i=\mu\}}\cdot 1_{\{h_j=\omega\}}\cdot 1_{\{y=\upsilon\}}\cdot\theta(y,h_i,h_j) \tag{7}$$

Similarly, atomic action and interaction activity have different pairwise features. For the atomic action, if $(i, j) \in E$, $f(h_i, h_j)$ is set to 1; otherwise it is set to 0. For complex activity, three pairwise features are extracted, including distance, intersection angle of moving orientations and difference of moving speeds between the two persons.

2.2 Two High-Level Features

The output of this modified-HCRF is two sets of high-level features conditioned on a certain action label rather than a decision made only by this one frame. The spatial constraints among local patches can be completely captured by the two features. The learned hidden state $\{h_1, h_2 \ldots h_N\}$ is used as the basic elements to generate these high-level representations due to its compact and semantic. The individual features and the interaction features are defined as follows.

Individual Features:

$$O^1_{\text{index}(\mu)}=\sum_{i\in V} p(\mathrm{h}_i=\mu\mid \mathrm{y},\mathbf{x},\theta)=\sum_{i\in V}\sum_{\mathrm{h}_i=\mu} p(\mathrm{h}\mid \mathrm{y},\mathbf{x},\theta)=\frac{\sum_{i\in V}\sum_{h_i=\mu} e^{\Phi(\mathrm{y},\mathrm{h},\mathbf{x};\theta)}}{\sum_{y\in Y,h\in H} e^{\Phi(\mathrm{y},\mathrm{h},\mathbf{x};\theta)}} \tag{8}$$

Interaction Features:

$$O^2_{\text{index}(\mu,\omega)}=\sum_{i,j\in E} p(\mathrm{h}_i=\mu,\mathrm{h}_j=\omega\mid \mathrm{y},\mathbf{x},\theta)=\sum_{i,j\in E}\sum_{\mathrm{h}_i=\mu,\mathrm{h}_j=\omega} p(\mathrm{h}\mid \mathrm{y},\mathbf{x},\theta)$$

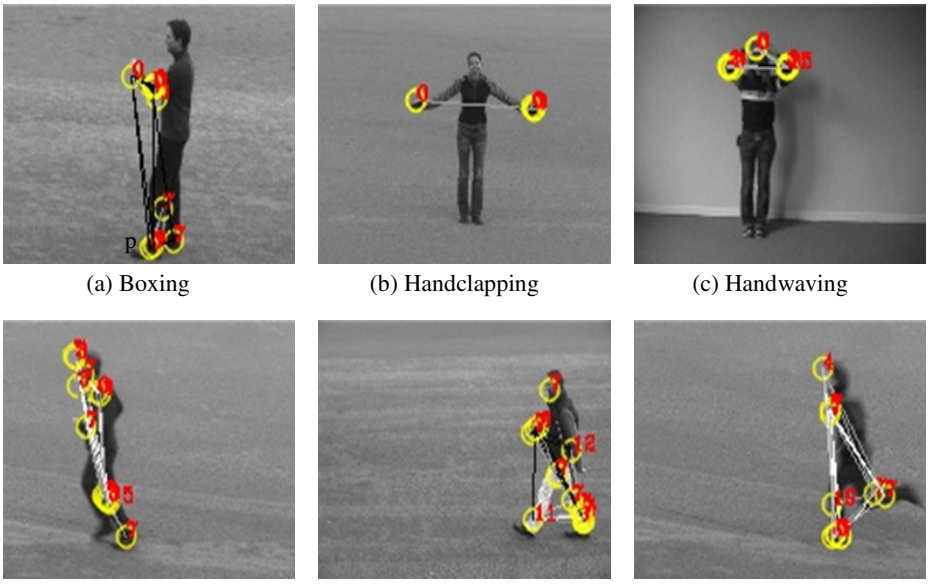
$$=\frac{\sum_{i,j\in E}\sum_{h_i=\mu,h_j=\omega} e^{\Phi(\mathrm{y},\mathrm{h},\mathbf{x};\theta)}}{\sum_{y\in Y,h\in H} e^{\Phi(\mathrm{y},\mathrm{h},\mathbf{x};\theta)}} \tag{9}$$

Where $\mu, \omega \in H$ refer to the hidden states. The $index(\mu)$ and $index(\mu, \omega)$ are the indexes of individual features and interaction features separately which are according to the hidden states. In Eq. (8) and Eq. (9), the two marginal probabilities $p(\mathrm{h}_i=\mu\mid \mathrm{y},\mathbf{x},\theta)$ and $p(\mathrm{h}_i=\mu,\mathrm{h}_j=\omega\mid \mathrm{y},\mathbf{x},\theta)$ can be calculated by belief propagation. No matter how the number of local patches or persons changes among different frames, dimension of the two features is fixed. It is only related to the number of hidden states.

Intuitively, the individual features are achieved by clustering similar features into a group. For example, to the action "walk", these individual features may be used to characterize the movement patterns of the left and right legs. The interaction features capture certain spatial constraints between pairs of these individual parts. In the case of "walk", two local patches at the left legs might have strong constraint that they tend to have the same state label, since both of them are characterized by the movement of the left leg. The two features cumulate the occurrence probability of each state and each interactional edge in a frame separately. Fig. 3 shows the visualization of the two high-level features in the KTH dataset.

3 Changeable Spatial-temporal Constraint Model

After feature learning, a new changeable spatial-temporal constraint model (CSTCM) is proposed for action classification which can make the utmost of these learned features. Our classification strategy specifies a local part filter and a changeable constraint model. The filter measures the similarity of appearance of local parts contained in a video sequence or a frame. The constraint model calculates the changeable cost of context for each local part. Because human actions appear in the three-dimensional space, both the dependencies in spatial and temporal domain should be concerned.



(a) Boxing (b) Handclapping (c) Handwaving

(d) Jogging (e) Running (f) Walking

Fig. 3. Images show the high-level individual features and interaction features learned from the modified-HCRF in the KTH action dataset. Each circle corresponds to a local individual feature with a red number which is the assignment of hidden states. Each line corresponds to an interaction feature between two individual features and different grayscales of the line represent different labels of interaction features.

For two video sequences v_i and v_j, the matching score $L(v_i, v_j)$ is the spatial score $S(v_i, v_j)$ of the whole video sequence plus the sum of temporal score $T(v_i^{t_1}, v_j^{t_2})$ over frames. Due to varying in time or speed, dynamic time warping (DTW) is used to measure their temporal similarity. Specially, T_i and T_j are frame numbers of the two videos which can be unequal in Eq. (10).

$$L(v_i, v_j) = S(v_i, v_j) + \sum_{\max(T_i, T_j)} \underset{t_1 \in T_i, t_2 \in T_j}{DTW} T(v_i^{t_1}, v_j^{t_2}) \tag{10}$$

Each of the scores includes two parts to describe the similarity of appearance and the change of constraint respectively. First, we define the spatial score between the two whole videos in Eq. (11).

$$S(v_i, v_j) = \sum_{\mu \in H} m_\mu(i, j) + \sum_{\mu, \omega \in H} d_{\mu,\omega}(i, j) \tag{11}$$

$$m_\mu(i, j) = (\sum_{t=1}^{T_i} O^1_{\text{index}(\mu_t^i)} - \sum_{t=1}^{T_j} O^1_{\text{index}(\mu_t^j)})^2 \tag{12}$$

$$d_{\mu,\omega}(i, j) = (\sum_{t=1}^{T_i} O^2_{\text{index}(\mu_t^i, \omega_t^i)} - \sum_{t=1}^{T_j} O^2_{\text{index}(\mu_t^j, \omega_t^j)})^2 \tag{13}$$

$m_\mu(i, j)$ is defined as a local part filter to measure the similarity of each individual features. It is something like the traditional BOF method, but we apply a feature learning algorithm to cluster local features. We all know that a local feature is not isolated but surrounded by its context. For example, when a local state occurs, another corresponding state always occurs as the same time. However, this relation may have some slight changes. Therefore, a constraint model $d_{\mu,\omega}(i, j)$ is used to calculate the changeable cost of interactional information for pair-wise local parts. $\sum_{\omega \in H} d_{\mu,\omega}(i, j)$ represents the whole cost of contextual change for a local part μ. Because only the spatial constraint is model here, the individual features and interaction features are cumulated over the whole video sequence.

Next, we define the temporal score with the same form that the first term denotes the local part filter and the second term refers to the constraint model. Unlike the spatial score, the temporal score is computed on each frame rather than the whole video to describe the temporal dependencies.

$$\begin{aligned} T(v_i^{t_1}, v_j^{t_2}) &= T_1(v_i^{t_1}, v_j^{t_2}) + T_2(v_i^{t_1}, v_j^{t_2}) \\ &= \sum_{\mu \in H} (O^1_{\text{index}(\mu_{t_1}^i)} - O^1_{\text{index}(\mu_{t_2}^j)})^2 + \sum_{\mu, \omega \in H} (O^2_{\text{index}(\mu_{t_1}^i, \omega_{t_1}^i)} - O^2_{\text{index}(\mu_{t_2}^j, \omega_{t_2}^j)})^2 \end{aligned} \tag{14}$$

There are three advantages in this classification model. Firstly, it takes the two key elements of human action into account, which are spatial and temporal constraints. Secondly, separating of the individual features from the interaction features in Eq. (12) and Eq. (13) makes the model easy to use the two high-level features independently and avoid them submerge each other. Through combining them in Eq. (11), the mutual influence of the two different features can be completely represented as well. Thirdly, rather than simple rigid matching, our CSTCM can tolerate some changes of local parts both on the spatial and temporal domain which is more adaptive to recognize action. Finally, the total matching score is employed in KNN for action matching.

4 Experiments

In this section, we evaluate our model both on the simple atomic action database and complex interaction activity database. To evaluate its effectiveness, we compare our algorithm with some other similar action recognition methods and some well-designed contrast experiments.

4.1 Databases

To test our proposed method, we use three well-known benchmark action recognition databases including both simple atomic action and complex interaction activity. For the simple atomic action, the public dataset KTH [12] and Weizmann [17] are used. The KTH contains 6 types of actions that are boxing, handwaving, running, handclapping, walking and jogging, under four different scenarios. Each video is captured at 25HZ with a size of 160×120. The Weizmann contains 93 video clips from 9 different actors. There are 10 different action categories: wave1, wave2, run, bend, walk, jump, side, jack, skip and pjump. Each video is captured at 25HZ with a size of 180×144. In these two datasets, each video contains only one actor performing a single action.

For the complex interaction activity, a two-person activity datasets CASIA [13] is adopted. It contains 84 video clips from three different views: angle view, horizontal view and top down view. There are 7 different activity categories: fight, followalways, followtogether, meetapart, meettogether, overtake and rob. So each activity has 12 clips and each view has 4 clips. The meaning of each activity is illustrated as follows and Fig.4 shows some example frames of two-person activities in the CASIA dataset.

Fight: Two persons are close to and then fight each other.

Followalways: One person follows another person and is behind him at a distance all the time.

Followtogether: One person follows another person and then they walk together.

Meetapart: Two persons are close to, then meet each other and depart at last.

Meettogether: Two persons are close to, then meet each other and walk together at last.

Overtake: One person is behind another person and then he overtakes.

Rob: One person follows another person, then robs him and runs away at last.

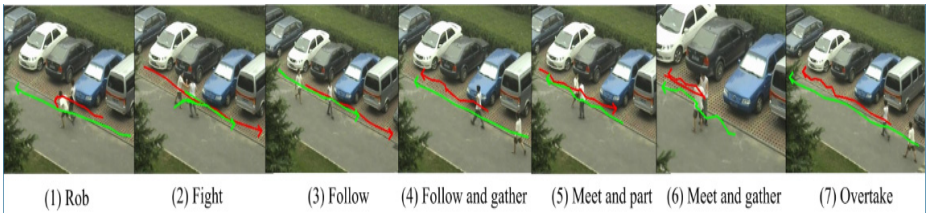


Fig. 4. Example frames of two-person activities in the CASIA dataset

4.2 Details of Our Model

For our model, the original input is the features directly extracted in videos. A 162 dimension vector including HOG (72) and HOF (90) around a local interested point is treated as the descriptor for the atomic action. For the interactional activity, five trajectory based features are applied. The size of possible hidden states in our modified-HCRF is 20, so we can learn 20 individual features and 400 interaction features. Then, these learned high-level features are applied in CSTCM for action classification, and the parameter K in KNN is set to 1 for simply.

A leave-one-out cross-validation strategy is used for Weizmann and CASIA. For KTH dataset, we use 16 actors for training and other 9 actors for testing like other papers' setup.

4.3 Recognition Results

Our proposal includes two parts: feature learning and CSTCM categorization. To evaluate the model completely, we test each of them separately.

Evaluate Feature Learning

Three contrast experiments are designed to evaluate our feature learning process, where the same descriptor is used as the original feature. Firstly, we classify every frame in a video sequence which is the standard HCRF method (per-frame recognition). Secondly, through the majority voting of the labels of its frames, the class label for the whole video sequence can be achieved (per-video recognition). Lastly, we use the two high-level features learned from the modified-HCRF to calculate histogram in the whole video sequence for action matching (histogram based method). Fig. 5 shows the visualization of the histogram based method in the CASIA dataset. Table 1 lists the recognition results of the three approaches on the KTH and CASIA dataset. It shows the following point:

- The accuracy of per-frame approach is lower than the per-video method. Because action is a continuous process, it is not enough to represent the whole action with only one frame. For example, the frame with the pose of "stand" may occur in all kinds of actions.

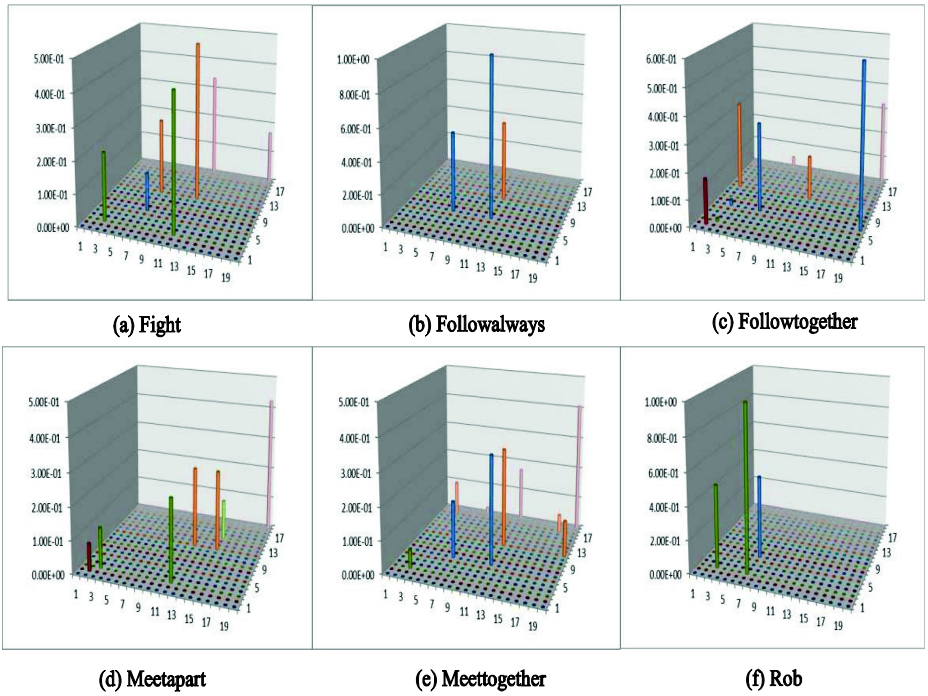


Fig. 5. Visualization of the histogram based method in the CASIA dataset. The histogram is cumulated by the high-level features learned from the whole video sequence. Elements on the primary diagonal are individual features and others denote the interaction features.

- The performance of the high-level features based approach (histogram based approach) is better than another two methods with low-level features (per-frame and per-video). It evidences that there is no much information lost and these learned features become more compact and semantic for action representation.

Table 1. A comparison of the three methods in feature learning part test on the KTH and CASIA dataset (%)

	Per-frame	Per-video	Histogram based
KTH	34.35	37.17	93.17
CASIA	58.98	84.52	91.67

Evaluate CSTCM

There are a local part filter and a changeable constraint model in the CSTCM. And both the *spatial and* temporal dependencies are modeled. For evaluating the model completely, we design some contrast experiments to test each of them separately. Table 2 lists their recognition results on the KTH, CASIA and the meaning of each variant is illustrated as follows.

Temporal score: the method only using temporal score to calculate the difference between two videos in the temporal domain.

Spatial score: the method only using *spatial* score to calculate the difference between two videos in the *spatial* domain and it is just the histogram based approach in Table 1.

Spatial score1: the method only using the local part filter in the *spatial* domain to measure similarity of appearance of local parts contained in the two whole video sequences. It is something like the traditional BOF method.

Spatial score2: the method only using the changeable constraint model in the *spatial* domain to calculate the changeable cost of context in the two whole video sequences.

CSTCM1: the method using the local part filter both in the *spatial*-temporal domain which can completely evaluate contribution of the local part filter.

CSTCM2: the method using the changeable constraint model both in the *spatial*-temporal domain which can completely evaluate contribution of the changeable constraint model.

CSTCM: the whole model proposed in this paper which contains the local part filter and changeable constraint model and both the differences in *spatial*-temporal domain are measured.

Table 2. A comparison of some variants with our proposal test on the KTH, CASIA dataset (%) and the mathematically definition of their matching scores

	Matching score ($L(v_i, \mathrm{v}_j) =$)	**KTH**	**CASIA**
Spatial score1	$\sum_{\mu \in H} m_\mu(i,j)$	82.33	89.29
Spatial score2	$\sum_{\mu,\omega \in H} d_{\mu,\omega}(i,j)$	92	90.48
Spatial score	$\mathrm{S}(v_i, \mathrm{v}_j)$	93.17	91.67
Temporal score	$\sum_{\max(\mathrm{T}_i,\mathrm{T}_j)} \underset{t_1 \in \mathrm{T}_i, t_2 \in \mathrm{T}_j}{DTW} T(v_i^{t_1}, \mathrm{v}_j^{t_2})$	78	45.24
CSTCM	$\mathrm{S}(v_i, \mathrm{v}_j) + \sum_{\max(\mathrm{T}_i,\mathrm{T}_j)} \underset{t_1 \in \mathrm{T}_i, t_2 \in \mathrm{T}_j}{DTW} T(v_i^{t_1}, \mathrm{v}_j^{t_2})$	96.17	92.86
CSTCM1	$\sum_{\mu \in H} m_\mu(i,j) + \sum_{\max(\mathrm{T}_i,\mathrm{T}_j)} \underset{t_1 \in \mathrm{T}_i, t_2 \in \mathrm{T}_j}{DTW} T_1(v_i^{t_1}, \mathrm{v}_j^{t_2})$	88.67	90.48
CSTCM2	$\sum_{\mu,\omega \in H} d_{\mu,\omega}(i,j) + + \sum_{\max(\mathrm{T}_i,\mathrm{T}_j)} \underset{t_1 \in \mathrm{T}_i, t_2 \in \mathrm{T}_j}{DTW} T_2(v_i^{t_1}, \mathrm{v}_j^{t_2})$	92.17	91.67

Note that when we evaluate contribution of the local part filter, only the individual features are used. While when the constraint model is test, only the interaction features are applied. From the table, it indicates following points.

- The performance of changeable constraint model is better than local part filter no matter in the spatial or temporal domain. Because changeable constraint model allows some non-rigid matching and the interaction features also contain more information for classification.
- The methods combining local part filter and changeable constraint model achieve a higher accuracy than using the two independently. Because when only using local part filter, the interactional constraint is lack and vice versa. It proves that each part can play a complementary role in most cases.
- No matter for the local part filter or changeable constraint model, when calculating both in the spatial and temporal domain, it can reach a better recognition result. It illustrates that human action has the characteristics of space and time. Both of them should be modeled.

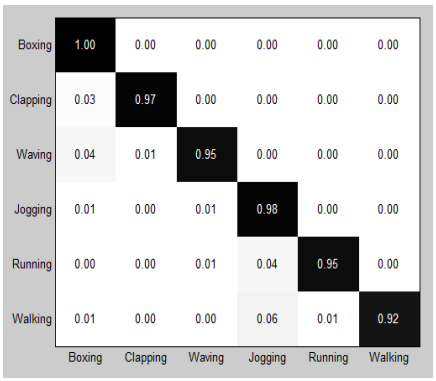
Compare with Other Methods

Finally, we compare our proposal with some other similar algorithms and it is summarized in Table 3. These similar approaches contain HCRF based methods (eg. [6], [7]); graphical model based methods for activity recognition (eg. [9], [14]); hand-designed contextual features methods (eg. [3], [15], [16], [19]); unsupervised feature learning approach [18]; and specially [2] is the traditional BOF. Note that the original local features and motion features adopted in our model are the same as [2, 15, 9], where it can fully evidence the effectiveness of our feature learning. From this table, we can find our recognition accuracy outperforms or is similar to all previous published results. Though the best result on Weizmann is just slightly higher than ours, the performance of our model on KTH and CASIA is much better. The results obtained by our proposed method on the KTH, Weizmann and CASIA dataset are reported in Fig.6 in detail which are the confusion matrixes for action classification.

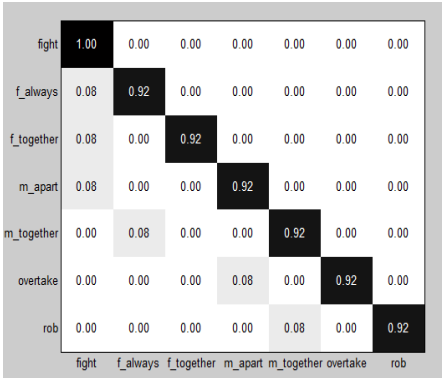
Table 3. A comparison of our proposal with other algorithms test on the KTH, Weizmann and CASIA dataset (%)

Algorithm	Weizmann	KTH	CASIA
Wang et al.[6]	97.22	87.6	—
Wang et al.[7]	100	92.51	—
Guo et al. [9]	—	—	83.28
Brand et al. [14]	—	—	76.14
Bregonzio al.[3]	96.6	93.17	—
Jiang et al. [15]	—	93.8	—
Wu et al. [16]	—	94.5	—
Fathi et al.[19]	100	90.5	—
Dollar et al. [2]	—	Over 80	—
Quoc et al. [18]	—	93.9	—
Our model	98.89	96.17	92.86

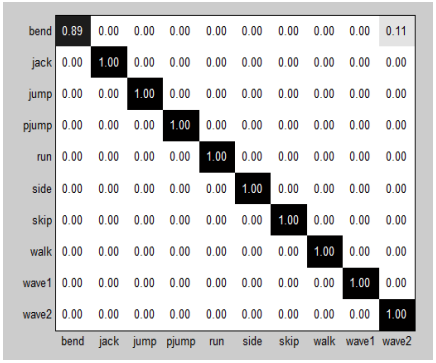
It is significantly inspired that through replacing hand-designed features by learned features and adopting a corresponding classification model CSTCM, we achieve a huge improvement both on simple atomic action database and complex interaction activity database. That is to say, our model not only proposes an effective feature leaning method, but also finds a way to make the best of what have been learned.



(a) The confusion matrice of KTH dataset



(b) The confusion matrice of CASIA dataset.



(c) The confusion matrice of Weizmann dataset.

Fig. 6. The confusion matrices for action classification on KTH (a), CASIA (b) and Weizmann (c)

5 Conclusions

The paper proposes a unified framework for human atomic action and interaction activity recognition. A supervised way is introduced to replace traditional hand-designed features by high-level learned features with the modified-HCRF. A new categorization model is used for corresponding classification which can make the utmost of the individual features and interaction features. The CSTCM specifies a local part filter and a changeable constraint model. In the spatial and temporal domain, the similarity of local

appearance is measured by the former, while the change of dependency is estimated by the latter. Therefore in this paper, actions can be modeled by local features in a changeable spatial-temporal constraint. Our method is evaluated on the KTH, Weizmann and CASIA dataset with significant improvement results.

Acknowledgements. This work is supported by the NSFC 61273274, 973 Program 2011CB302203, National Key Technology R&D Program of China 2012BAH01F03, NSFB4123104, Z131110001913143, Tsinghua-Tencent Joint Lab for IIT and Beijing Jiaotong University Research Foundation Program KKJB14029536.

References

1. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64**(2–3), 107–123 (2005)
2. Dollár, P., et al.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. IEEE (2005)
3. Bregonzio, M., Gong, S., Xiang, T.: Recognising action as clouds of space-time interest points. In: CVPR (2009)
4. Ryoo, M., Aggarwal, J.: Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In: ICCV (2009)
5. Sun, J., Wu, X., Yan, S., Cheong, L.F., Chua, T.-S., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: CVPR (2009)
6. Wang, Y., Mori, G.: Learning a discriminative hidden part model for human action recognition. In: NIPS, vol. 8 (2008)
7. Wang, Y., Mori, G.: Max-margin hidden conditional random fields for human action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009. IEEE (2009)
8. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008, CVPR 2008. IEEE (2008)
9. Guo, P., et al.: Coupled Observation Decomposed Hidden Markov Model for Multiperson Activity Recognition. IEEE Transactions on Circuits and Systems for Video Technology **22**(9), 1306–1320 (2012)
10. Quattoni, A., Collins, M. , Darrell, T.: Conditional random fields for object recognition (2004)
11. Wang, S.B., et al.: Hidden conditional random fields for gesture recognition. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2. IEEE (2006)
12. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: ICPR (2004)
13. CASIA Action Database. http://www.cbsr.ia.ac.cn/english/Action%20Databases%20EN.asp
14. Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 994–999 (1997)
15. Wang, J., Chen, Z., Wu, Y.: Action recognition with multiscale spatio-temporal contexts. In: CVPR (2011)
16. Wu, X., Xu, D., Duan, L., Luo, J.: Action recognition using context and appearance distribution features. In: CVPR (2011)
17. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV (2005)

18. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR (2011)
19. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: CVPR (2008)
20. Vahdat, A., Cannons, K., Mori, G., et al.: Compositional models for video event detection: a multiple kernel learning latent variable approach [C]. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 1185–1192. IEEE (2013)