# Human Action Recognition Based on Context-Dependent Graph Kernels

Baoxin Wu, Chunfeng Yuan, and Weiming Hu
National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
{bxwu,cfyuan,wmhu}@nlpr.ia.ac.cn

## Abstract

*Graphs are a powerful tool to model structured objects, but it is nontrivial to measure the similarity between two graphs. In this paper, we construct a two-graph model to represent human actions by recording the spatial and temporal relationships among local features. We also propose a novel family of context-dependent graph kernels (CGKs) to measure similarity between graphs. First, local features are used as the vertices of the two-graph model and the relationships among local features in the intra-frames and inter-frames are characterized by the edges. Then, the proposed CGKs are applied to measure the similarity between actions represented by the two-graph model. Graphs can be decomposed into numbers of primary walk groups with different walk lengths and our CGKs are based on the context-dependent primary walk group matching. Taking advantage of the context information makes the correctly matched primary walk groups dominate in the CGKs and improves the performance of similarity measurement between graphs. Finally, a generalized multiple kernel learning algorithm with a proposed $l12$-norm regularization is applied to combine these CGKs optimally together and simultaneously train a set of action classifiers. We conduct a series of experiments on several public action datasets. Our approach achieves a comparable performance to the state-of-the-art approaches, which demonstrates the effectiveness of the two-graph model and the CGKs in recognizing human actions.*

## 1. Introduction

Many of successful methods for human action recognition are based on local spatio-temporal features [4, 17, 11, 22], which are extracted sparsely from video sequences. In these methods, an action is represented as an ensemble of local features. This ensemble contains not only individual local features but also complex topological structure among these local features. Graphs are an effective tool for modeling complex structured data [1, 8]. However, few researches model the ensemble of local features by graphs in human action recognition. There are two nontrivial difficulties to be solved: i) how to construct graphs to model these local features; ii) how to measure similarity between the constructed graphs. In this paper, we focus on these two problems and propose a new graph-based approach for human action recognition.

To model the ensemble of local features, we construct two directed and attributed graphs, based on the local features with intra-frame relationships and inter-frame relationships. These two graphs are named as *video co-occurrence graph* (VCG) and *video successiveness graph* (VSG) respectively. The vertex attributes in both graphs correspond to the local features of a video sequence. The edge attributes in VCG and VSG describe the spatial layout relationships of local features detected in the intra-frames and in the inter-frames respectively. The VCG and VSG are complementary to each other. Compared with the popular bag-of-words model [4, 22], these two graphs preserve not only the individual power of local features but also most of the spatio-temporal relationships among them, and hence they provide a more informative representation for actions.

As the actions represented by VCGs and VSGs are structured, it is difficult to directly use the traditional statistical classification methods to classify them. We propose a novel family of *context-dependent graph kernels* (CGKs) which are a bridge between the structured action representation and the statistical classification. The proposed CGKs are actually a series of decomposition kernels. Specifically, graphs are first decomposed into a number of primary walk groups (PWGs) with different lengths, and CGKs are then obtained by context-dependent matching of PWGs from two graphs. The main property of CGKs is that the contexts of PWGs are incorporated into the PWGs matching, which improves the similarity measurement between PWGs. Usually, only the correctly matched PWGs carry meaningful discriminant information for comparing graphs. The correctly matched PWGs should have high similarity value not only between themselves, but also between their contexts. In CGKs, the incorporation with context information makes the correctly matched PWGs dominate and improves the performance of similarity measurement between

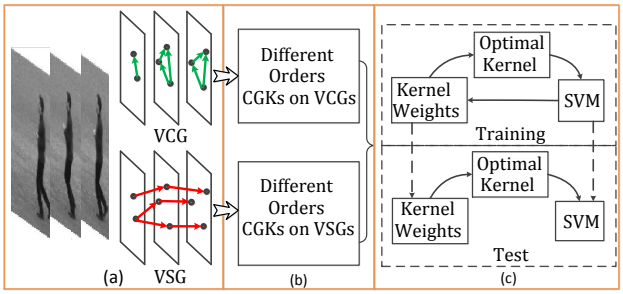


Figure 1. The illustration of the CGKs based human action recognition. (a) A video sequence is represented by VCG and VSG together. (b) Different orders CGKs are computed on both video graphs. (c) The GMKL algorithm is applied to combine the CGKs together and learn action classifiers simultaneously.

graphs. However, in the traditional random walk graph kernels (TGKs) [5, 1, 20], where all the possible matchings between PWGs are summed with same weights, the discriminative power of correctly matched PWGs is swamped by that of the incorrectly matched ones.

The proposed CGKs can efficiently measure the similarity between graphs. Subsequently, a generalized multiple kernel learning (GMKL) algorithm with $l12$-*norm regularization* is applied to combine together different order CGKs on both graphs. The proposed $l12$-norm regularization considers both the sparseness constraint on the kernels from the same graph and the smoothness constraint on kernels from different graphs. The logical block diagram of our approach is shown in Figure 1.

The main contributions of this paper are as follows:

- A two-graph model is proposed for human action representation, capturing the spatio-temporal relationships among local features.
- A novel family of CGKs is proposed to measure similarity between attributed graphs. The CGKs utilize the contexts of PWGs to improve the performance of similarity measurement between graphs.
- A GMKL formulation with $l12$-norm regularization is applied for kernel combination and action classification.

## 2. Related Work

Graphs are a natural tool for modeling structured data, with vertices representing parts and edges representing the relations between them. They have been widely applied and shown good performance in the fields of protein prediction and chemical molecular analysis [1, 7].

Graphs have also been utilized in human action recognition. Borzeshi *et al.* [2] represent each frame as a graph with vertices corresponding to the spatial local features extracted from this frame. Raja *et al.* [14] describe a person in a frame with a graphical model which contains six vertices encoding the positions of five body parts and the action label. Gaur *et al.* [6] construct a string of feature graphs for the spatiotemporal layout of local features. Each graph in the string models the spatial configuration of local features in a small temporal segments. Ta *et al.* [18] construct a hypergraph to model the extracted spatiotemporal local features and a hypergraph matching algorithm is used for activity recognition. These approaches construct graphs to model local features or body parts. However, they do not provide explicitly spatio-temporal relationships between these local features or body parts.

For the similarity measurement between graphs, random walk graph kernels have received increasing attention recently. Gärtner *et al.* [5] compute the graph kernel on two labeled graphs by counting the number of matched labeled random walks. Then it is extended by Borgwardt *et al.* [1] by replacing the Dirac kernel with more complex kernels for continuous attriubtes. Vishwanathan *et al.* [20] propose a generalized version of the random walk graph kernels and introduce several techniques to speed up the computation of random walk graph kernels. Harchaoui and Bach [8] build a set of segmentation graph kernels on images and utilize a multiple kernel learning method to combine these kernels together and classify images. These graph kernels are all built by comparing the similarities between all pairs of walks from two graphs. However, the contexts of walks, which can improve the similarity measurement between walks, are not exploited in all of their approaches.

## 3. Context-Dependent Graph Kernels

### 3.1. Construction of CGKs

We propose a family of CGKs for the similarity measurement between attributed graphs. Graphs are first decomposed into a number of PWGs and CGKs are obtained by the PWG matching, incorporating with the context information of PWGs.

An attributed graph with $N$ vertices is denoted as $G = (V, E)$, where $V = \{v_i\}_{i=1}^{N}$ is the vertex set and $E$ is the edge set. A vertex is a point embedded in a Euclidean space with a vector of attributes attached to it. It is defined as $v_i = (l_i, d_i)$, where $l_i \in R^u$ is the coordinate of the vertex and $d_i \in R^z$ is the corresponding attribute vector. In the paper, $u = 3$ and $z$ is the dimension of local feature descriptors. If a vertex $v_j$ is a neighbor of $v_i$, $v_i$ and $v_j$ form an edge $(v_i, v_j) \in E$.

Our graph decomposition is based on random walks. A random walk with length $n$ from graph $G$ is denoted as a sequence of vertices jointed by edges, $w = (v_{w_0}, e_{w_1}, ..., e_{w_n}, v_{w_n})$, where $e_{w_i} = (v_{w_{i-1}}, v_{w_i}) \in E$, $1 \leq i \leq n$. Let $\rho_G^n$ be the set of total walks with length $n$

in graph G and $\rho_G^n(i,j) \subset \rho_G^n$ be a primary walk group (PWG), namely, a subset of $\rho_G^n$ containing walks starting at vertex $v_i$ and ending at $v_j$. It means that when a walk $w \in \rho_G^n(i,j)$, we have $v_{w_0} = v_i$ and $v_{w_n} = v_j$. Actually, for a walk $w$ with length $n = 0$, the walk is a vertex. Therefore, we have $\rho_G^0 = V$, and $\rho_G^0(i,i) = v_i$. The PWGs can be regarded as substructures of graphs and the context-dependent kernels on PWGs constitute the CGKs on graphs.

Let $k_v(v,v')$ and $k_e(e,e')$ be two kernel functions defined on vertices and edges respectively. These two functions can be designed differently according to different tasks. Let $k_w(w,w')$ be a kernel function on two walks with the same length $n$. If $n = 0$, we have $k_w(w,w') = k_v(v_w, v'_{w'})$. If $n \geq 1$, we have

$$k_w(w,w') = \prod_{i=0}^{n} k_v(v_{w_i}, v'_{w'_i}) \prod_{j=1}^{n} k_e(e_{w_j}, e'_{w'_j}). \quad (1)$$

The kernels on PWGs are defined as a summation of walk kernels on all pairs of walks from both PWGs

$$k_{wg}(\rho_G^n(i,j), \rho_G^n(r,s)) = \sum_{w \in \rho_G^n(i,j)} \sum_{w' \in \rho_G^n(r,s)} k_w(w,w'). \quad (2)$$

Subsequently, we define the contexts of a PWG $\rho_G^n(i,j)$ as $\eta_G^n(i,j)$, which has the following form

$$\eta_G^n(i,j) = \{\rho_G^n(p,q) |\, p \in c(i), q \in c(j)\}, \quad (3)$$

where $c(i)$ denotes the contexts of vertex $v_i$. We define $c(i)$ as the set of the $m$ nearest vertices of $v_i$ in the Euclidean space, formulated as follows

$$c(i) = \{v_p |\, ||l_i - l_p||_2 \leq ||l_i - l_q||_2, \forall\, v_q \notin c(i) \text{ and } |c(i)| = m, p \neq i\}. \quad (4)$$

The similarity between the contexts of PWGs is used as a weight on the similarity between the PWGs. The context-dependent kernels on PWGs are defined as

$$k_{cwg}(\rho_G^n(i,j), \rho_G^n(r,s)) = k_{wg}(\rho_G^n(i,j), \rho_G^n(r,s)) * (1 + \kappa k_{wg}(\eta_G^n(i,j), \eta_G^n(r,s))), \quad (5)$$

where

$$k_{wg}(\eta_G^n(i,j), \eta_G^n(r,s)) = \sum_{\substack{p \in c(i),\, q \in c(j),\\ k \in c(r),\, t \in c(s)}} k_{wg}(\rho_G^n(p,q), \rho_G^n(k,t)) \quad (6)$$

and $\kappa$ is a constant controlling the weight of the context information in the kernel. Figure 2 shows the kernels on PWGs and on their contexts.

The proposed CGKs on graphs are computed as a summation of context-dependent similarities between all pairs of PWGs from two graphs. Given two graphs $G = (V,E)$ and $G' = (V',E')$, we refer to the CGK with respect to walk length $n$ as the $n$th-order CGK, which is defined as

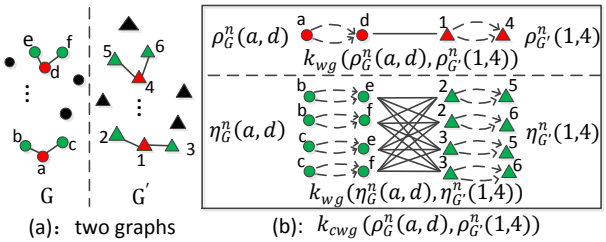


Figure 2. (a) shows two graphs $G$ and $G'$. The green vertices connected with the red ones are their contexts. The edges are not shown in both graphs. (b) shows the kernels on PWGs and the kernels on their contexts. These two parts constitute the context-dependent kernel on PWGs by equation 5.

$$k_g^n(G,G') = \frac{1}{N_G^n N_{G'}^n} \sum_{\substack{\rho_G^n(i,j) \subset \rho_G^n\\ \rho_{G'}^n(r,s) \subset \rho_{G'}^n}} k_{cwg}(\rho_G^n(i,j), \rho_{G'}^n(r,s)) \quad (7)$$

where $N_G^n$ and $N_{G'}^n$ are the numbers of PWGs with length $n$ in $G$ and $G'$ respectively. With a sequence of weight variable $(\lambda_0, \lambda_1, \lambda_2, ...)$ to emphasize the importance of each order $k_g^n(G,G')$, the final graph kernel is computed as a weighted summation of the different $n$th-order CGKs

$$k_g(G,G') = \sum_{n=0} \lambda_n k_g^n(G,G'). \quad (8)$$

### 3.2. Relations to Other Kernels

CGKs are related to the traditional random walk graph kernels (TGKs). TGKs are computed as a summation of similarities between all pairs of walks from two graphs, i.e., the summation of similarities between all pairs of PWGs. The $n$th-order TGK is expressed as

$$k_{tg}^n(G,G') = \sum_{w \in \rho_G^n} \sum_{w' \in \rho_{G'}^n} k_w(w,w') = \sum_{\substack{\rho_G^n(i,j) \subset \rho_G^n,\\ \rho_{G'}^n(r,s) \subset \rho_{G'}^n}} k_{wg}(\rho_G^n(i,j), \rho_{G'}^n(r,s)). \quad (9)$$

Except for the normalization factor, TGKs can be viewed as a special case of our CGKs where $\kappa$ in equation 5 is zero. It means that the context information of PWGs is not utilized in the matching of PWGs. All pairs of matched PWGs are combined with the same weights. However, this can reduce the discriminative power of correctly matched PWGs from two graphs.

As the walk with length $n = 0$ is actually a vertex, the 0th-order CGK can be rewritten equally as

$$k_g^0(G,G') = \frac{1}{|V||V'|} \sum_{i,j} k_v(v_i, v'_j)(1 + \kappa \sum_{\substack{k \in c(i),\\ t \in c(j)}} k_v(v_k, v'_t)). \quad (10)$$

It is determined only by the vertex information of graphs and no edge information is considered in this kernel. When $\kappa$ is large enough, it is an approximation of the neighborhood kernel on the attributed point sets [13]

$$k_{nk}(V,V') = \frac{1}{|V||V'|}\sum_{i,j} k_v(v_i,v'_j)\sum_{\substack{k\in c(i),\\ t\in c(j)}} k_v(v_k,v'_t). \quad (11)$$

Therefore, the CGKs are an extension of the neighborhood kernel from attributed point sets to attributed graphs.

### 3.3. Computation of CGKs

For two graphs $G$ and $G'$, it has been proved that performing a random walk on the direct product graph of these two graphs is equivalent to performing a simultaneous random walk on $G$ and $G'$ [9]. When computing the CGKs in practice, we utilize the direct product graph to make the computation efficient. The direct product graph of $G$ and $G'$ is denoted as $G_P = (V_P, E_P)$, where $V_P$ and $E_P$ are defined as

$$V_P = \{(v_i, v'_r) \mid v_i \in V, v'_r \in V', k_v(v_i, v'_r) > 0\} \quad (12)$$

and

$$E_P = \{((v_i,v'_r),(v_j,v'_s)) \mid (v_i,v_j)\in E, (v'_r,v'_s)\in E', \\ k_e((v_i,v_j),(v'_r,v'_s)) > 0\} \quad (13)$$

respectively. For each vertex $(v_i, v'_r) \in V_P$, we assign it a weight $\omega_{ir} = k_v(v_i, v'_r)$ and for each edge $((v_i,v'_r),(v_j,v'_s)) \in E_P$, we assign it a weight $\omega_{ir,js} = k_e((v_i,v_j),(v'_r,v'_s))$, where $ir$ and $js$ are the vertex indices in $G_P$. We use two matrixes $W_V$, $W_E \in R^{|V_P|\times|V_P|}$ to contain the vertex weights and edge weights, with $[W_V]_{ir,ir} = \omega_{ir}$ and $[W_E]_{ir,js} = \omega_{ir,js}$ respectively. The final $n$th-order context-free weight matrix $W_P^n$ of $G_P$ is expressed as

$$W_P^n = W_V(W_E W_V)^n, \quad (14)$$

We define the context matrix $C_P \in R^{|V_P|\times|V_P|}$ for vertices in $G_P$ as

$$[C_P]_{ir,js} = \begin{cases} 1 \text{ if } j \in c(i) \text{ and } s \in c(r) \\ 0 \text{ otherwise} \end{cases} \quad (15)$$

and the $n$th-order context-dependent weight matrix $W_{CP}^n$ of $G_P$ is

$$W_{CP}^n = W_P^n + \kappa W_P^n \odot (C_P W_P^n C_P^T), \quad (16)$$

where $\odot$ represents the Hadamard product which is obtained by the element-wise multiplication of two matrices with the same size and $C_P^T$ is the transposed matrix of $C_P$.

In fact, the elements of $W_P^n$ and $W_{CP}^n$ correspond to the context-free and context-dependent kernels on PWGs respectively. So we have

$$[W_P^n]_{ir,js} = k_{wg}(\rho_G^n(i,j), \rho_G^n(r,s)), \quad (17)$$

and

$$[W_{CP}^n]_{ir,js} = k_{cwg}(\rho_G^n(i,j), \rho_G^n(r,s)). \quad (18)$$

According to equation 7, the $n$th-order CGK on graphs is expressed as

$$k_g^n(G,G') = \frac{1}{N_G^n N_{G'}^n}\sum_{ir,js}[W_{CP}^n]_{ir,js} \quad (19)$$

Substituting the above equation into equation 8, we obtain the final graph kernel on $G$ and $G'$.

## 4. CGKs Based Action Recognition

### 4.1. Two-Graph Model for Action Representation

Given a video sequence, we first extract its local spatio-temporal features. We utilize Dollár's separable linear filters [4] to detect the spatio-temporal interest points in the video sequence and the 3D SIFT descriptor [17] to describe the obtained interest points. Let $N$ be the number of total interest points and $f_i = [l_i, d_i]$ be the $i$th local feature where $l_i = (x_i, y_i, t_i)$ is the space-time coordinate in the 3D domain and $d_i$ denotes the 3D SIFT descriptor. So the ensemble of local features is depicted as $\{f_1, f_2, ..., f_N\}$.

We construct two graphs: a video co-occurrence graph (VCG) and a video successiveness graph (VSG), to model the spatial layout relationships of local features detected in the intra-frames and inter-frames. These two graphs themselves reflect different temporal order relationships between local features.

The VCG and VSG are denoted as $G_c = (V_c, E_c, A_c)$ and $G_s = (V_s, E_s, A_s)$ respectively, where $V_c$ and $V_s$ are the vertex sets, $E_c$ and $E_s$ are the edge sets, $A_c$ and $A_s \in R^{N\times N}$ are the affinity matrixes of the two graphs. The vertices of the two graphs correspond to interest points with their 3D SIFT descriptors as the vectors of attributes. We employ the $\varepsilon$-Graph method to construct $G_c$ and $G_s$. For $G_c$, $A_c$ has the form of

$$A_c(i,j) = \begin{cases} 1 & \text{if } \sqrt{(x_j-x_i)^2+(y_j-y_i)^2} < \varepsilon_1, \\ & \quad y_j \geq y_i, \text{and } t_j = t_i, \\ 0 & \text{otherwise.} \end{cases} \quad (20)$$

For $G_s$, $A_s$ has the form of

$$A_s(i,j) = \begin{cases} 1 & \text{if } \sqrt{(x_j-x_i)^2+(y_j-y_i)^2} < \varepsilon_2, \\ & \quad \text{and } \quad 0 < t_j - t_i \leq \varepsilon_t, \\ 0 & \text{otherwise.} \end{cases} \quad (21)$$

The parameters $\varepsilon_1$ and $\varepsilon_2$ are two thresholds of the spatial distance between two vertices and $\varepsilon_t$ is the threshold of the temporal distance. When $A_c(i,j) = 1$, we have $(v_i, v_j) \in E_c$ and when $A_s(i,j) = 1$, $(v_i, v_j) \in E_s$. Moreover, discrete attributes are attached to edges in both graphs according to the relative spatial positions of vertices. A polar coordinate system with the origin at the coordinate of $v_i$ is utilized to capture the relative spatial location information between $v_i$ and its neighbors. The polar coordinate is divided into a number of bins. The edge attribute of $(v_i, v_j)$ is represented by the index of the bin where $v_j$ locates on the polar coordinate system.

Actions are represented by VCGs and VSGs together. This representation has two properties. First, these two graphs are complementary to each other and preserve the spatial and temporal relationships between local features. Second, these two graphs are indeed directed graphs and there are no cycles in both of them. The random walks in both graphs are paths, which avoid the tottering and halting problems when computing graph kernels based on random walks.

### 4.2. Action Similarity Measurement by CGKs

We apply the proposed CGKs to measure the similarity between human actions represented by VCGs and VSGs. First of all, we define the vertex kernel as

$$k_v(v, v') = \begin{cases} \exp(-\frac{||d-d'||_2^2}{2\sigma^2}) & \text{if } ||d - d'||_2 \leq \varepsilon_d, \\ 0 & \text{otherwise,} \end{cases} \quad (22)$$

where $d$ and $d'$ are the corresponding 3D SIFT descriptors for vertices $v$ and $v'$ respectively, $\sigma$ is a scale parameter for Gaussian function, and $\varepsilon_d$ is a threshold. Meanwhile, we also define the edge kernel as

$$k_e(e, e') = \begin{cases} 1 & \text{if } attribute(e) = attribute(e'), \\ 0 & \text{otherwise.} \end{cases} \quad (23)$$

The above definitions of vertex and edge kernels reduce the size of the direct product graph and make it quite sparse, which speeds up the computation process.

Let $S = (G_c, G_s)$ and $S' = (G'_c, G'_s)$ be two video sequences. When computing the 0th-order CGK on two graphs, we have $k_g^0(G_c, G'_c) = k_g^0(G_s, G'_s)$, as the two graphs have the same vertex set. We set the maximal order of CGKs on VCGs and VSGs are $m_c$ and $m_s$ respectively. According to equation 8, the final kernel on two sequences is expressed as

$$\begin{aligned} k(S, S') =& \lambda_0 k_g^0(G_c, G'_c) + \sum_{i=1}^{m_c} \lambda_{c_i} k_g^i(G_c, G'_c) \\ &+ \sum_{j=1}^{m_s} \lambda_{s_j} k_g^j(G_s, G'_s). \end{aligned} \quad (24)$$

where $\Lambda = [\lambda_0, \lambda_{c_1}, ..., \lambda_{c_{m_c}}, \lambda_{s_1}, ..., \lambda_{s_{m_s}}]^T$, $\Lambda > 0$ is a weight vector and $k_g^0(G_c, G'_c), k_g^i(G_c, G'_c), k_g^j(G_s, G'_s)$ are computed by equation 19.

### 4.3. Generalized Multiple Kernel Learning

We apply the GMKL formulation proposed by Varma *et al.* [19] to learn the weight for each CGK and the action classifiers simultaneously. Assume a set of $M$ training sequences $\{S_p, y_p\}_{p=1}^M$ where $S_p = (G_{cp}, G_{sp})$ represents an input video sequence and $y_p$ is the action label associated with $S_p$. We define a set of $M \times M$ kernel matrixes $\{K_0, K_{c_1}, ..., K_{c_{m_c}}, K_{s_1}, ..., K_{s_{m_s}}\}$ for the training video sequences, with $[K_0]_{p,q} = k_g^0(G_{cp}, G_{cq})$, $[K_{c_i}]_{p,q} = k_g^i(G_{cp}, G_{cq})$, and $[K_{s_j}]_{p,q} = k_g^j(G_{sp}, G_{sq})$, where $1 \leq i \leq m_c$ and $1 \leq j \leq m_s$. According to equation 24, the final kernel matrix $K$ for the training sequences is defined as

$$K = \lambda_0 K_0 + \sum_{i=1}^{m_c} \lambda_{c_i} K_{c_i} + \sum_{j=1}^{m_s} \lambda_{s_j} K_{s_j}, \quad (25)$$

where $\Lambda = [\lambda_0, \lambda_{c_1}, ..., \lambda_{c_{m_c}}, \lambda_{s_1}, ..., \lambda_{s_{m_s}}]$ is the same as the kernel weight vector defined in equation 24.

The regularization of weights on CGKs is determined as follows. On the one hand, different $n$th-order $(n \geq 1)$ CGKs on the same graph may contain redundant information for recognizing actions. There should be a sparseness constraint on the weights of those kernels. On the other hand, as VCGs and VSGs preserve different spatio-temporal relationships between local features, CGKs on different graphs contain complementary information for action recognition. For the 0th-order CGK, the spatio-temporal relationships of local features are not considered in the kernel computation. So, there should be a smoothness constraint on the weights of CGKs from different graphs and the 0th-order CGK. Considering the above two aspects, we apply the GMKL framework with a *l*12-*norm* regularization on the kernel weights to join all the CGKs together optimally. The $l12$-norm regularization is modeled as

$$\begin{aligned} r(\Lambda) =& \frac{1}{2}|| \, |\lambda_0|_1, |\lambda_{c_1}, ..., \lambda_{c_{m_c}}|_1, |\lambda_{s_1}, ..., \lambda_{s_{m_s}}|_1 \, ||_2^2 \\ =& \frac{1}{2}(\lambda_0^2 + (\sum_{i=1}^{m_c} \lambda_{c_i})^2 + (\sum_{j=1}^{m_s} \lambda_{s_j})^2). \end{aligned} \quad (26)$$

Let $Y$ be a diagonal matrix with the action class labels $y_i$ on the diagonal and $K$ be the kernel matrix defined in equation 25. The dual problem of GMKL is represented as

$$\begin{aligned} &\min_{\Lambda} D(\Lambda) \\ &\text{where } D(\Lambda) = \max_{\boldsymbol{\alpha}} \mathbf{1}^T \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T Y K Y \boldsymbol{\alpha} + C_2 r(\Lambda) \\ &\text{subject to} \quad \mathbf{1}^T Y \boldsymbol{\alpha} = 0, \; 0 \leq \boldsymbol{\alpha} \leq C_1, \; \Lambda \geq 0, \end{aligned} \quad (27)$$

where $\boldsymbol{\alpha}$ is the Lagrangian multiplier and $C_1$, $C_2$ are two constants controlling the importance of hinge loss and regularization on kernel weights respectively.

As the $D(\Lambda)$ is differentiable, the derivatives of $D(\Lambda)$ have the form as

$$\begin{aligned}\frac{\partial D}{\partial \lambda_0} &= C_2\lambda_0 - \frac{1}{2}\boldsymbol{\alpha}^T Y K_0 Y \boldsymbol{\alpha},\\ \frac{\partial D}{\partial \lambda_{c_i}} &= C_2\sum_{k=1}^{m_c}\lambda_{c_k} - \frac{1}{2}\boldsymbol{\alpha}^T Y K_{c_i} Y \boldsymbol{\alpha},\\ \frac{\partial D}{\partial \lambda_{s_j}} &= C_2\sum_{k=1}^{m_s}\lambda_{s_k} - \frac{1}{2}\boldsymbol{\alpha}^T Y K_{s_j} Y \boldsymbol{\alpha},\end{aligned} \tag{28}$$

where $1 \le i \le m_c$ and $1 \le j \le m_s$.

Then the minimax optimization algorithm is utilized to calculate $\boldsymbol{\alpha}$ and $\Lambda$ iteratively. In the first stage, $\Lambda$ is fixed so that $K$ and $r(\Lambda)$ are constants. In this situation, $\boldsymbol{\alpha}$ can be obtained by applying any SVM solver. In the second stage, $\boldsymbol{\alpha}$ is kept fixed and $\Lambda$ is estimated by the projected gradient descent method. The weights are updated by $\lambda_k^{t+1} = \lambda_k^t - s^t(\partial D/\partial\lambda_k)$ and projected to a feasible set $\lambda_k^{t+1} = \max(0, \lambda_k^{t+1})$, where $k \in \{0, c_1, ..., c_{m_c}, s_1, ..., s_{m_s}\}$ and $s^t$ is the step size chosen based on the Armijo rule. These two stages are repeated iteratively until convergence.

# 5. Experimental Results

We test the proposed action recognition approach on several benchmark datasets: the KTH action dataset [16], the UCF Sports dataset [15], and the UCF Films dataset [15].

Six approaches are designed in order to evaluate and contrast with the performance of our algorithm in recognizing human actions. In the first approach, we use a bag-of-words model to represent the ensemble of local features extracted from a video sequence. A $\chi^2$-kernel is used to measure the similarity between the histograms of sequences and an SVM classifier is used for action classification. Obviously, the spatio-temporal relationships of local features are not involved in this approach.

In the second approach, we evaluate the performance of the 0th-order CGK. Though video sequences are represented by VCGs and VSGs, the edge information of two graphs are not considered in this kernel. It depends on the individual discriminative power of the local features, involving no spatio-temporal relationships among them. The obtained kernel matrix is sent into an SVM classifier directly for human action classification.

In the third approach, the ensemble of local features is modeled by the VCG. We compute different order CGKs (including the 0th-order CGK) on VCGs and apply the GMKL to join the obtained kernels together. In the fourth approach, the ensemble of local features is modeled by the VSG. We also compute different order CGKs (including the 0th-order CGK) on VSGs and apply the GMKL to join the kernels together. In these two approaches, sequences are represented by different graphs, which exploit different spatio-temporal relationships among local features.

In the fifth and sixth approaches, the ensemble of local features is modeled by the VCG and VSG together. In the fifth approach, traditional random walk graph kernels (TGKs) are used to measure the similarity between video graphs and different order TGKs on both graphs are combined by the GMKL algorithm. In the sixth approach, we use our proposed CGKs to measure the similarity between video graphs and combine the obtained CGKs together by the GMKL algorithm. The main difference between these two approaches is that the contexts of PWGs play a role in the PWGs matching when computing CGKs.

The above six approaches are referred to as 'BoW', '0-CGK', 'VCG+CGKs','VSG+CGKs','VCG+VSG+TGKs' and 'VCG+VSG+CGKs' respectively. In our experiments, we define the nearest 5 interest points in the 3D space as the contexts of a given interest point. We set the max order of CGKs on VCGs and VSGs to be 4 and 5 respectively.

## 5.1. Experiments on the KTH Dataset

The KTH dataset is a widely used action dataset which contains six human action classes. They are performed by 25 subjects under four different scenarios. There are total 599 video sequences in this dataset. We perform the leave-one-out cross validation, i.e., videos of 24 subjects are used for training and videos of the remaining one subject are used for test. We perform the above mentioned six approaches on the KTH dataset. For the 'BoW' approach, the number of visual words in the vocabulary is set to 1600 according to the cross validation. The average accuracy values on all classes of the six approaches are shown in Table 1. The performances of six approaches on each action class are shown in Figure 3.

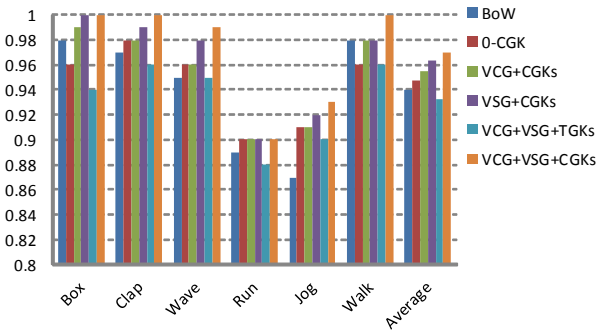


Figure 3. The performance of six approaches on the KTH dataset.

Table 1 and Figure 3 show that the approach of 'VCG +VSG+CGKs' achieves the best accuracy value 97.0% on the KTH dataset. Moreover, the following four points can

| | KTH (%) | UCF (%) |
|---|---|---|
| BoW | 93.9 | 84.7 |
| 0-CGK | 94.7 | 85.3 |
| VCG+CGKs | 95.5 | 88.7 |
| VSG+CGKs | 96.3 | 89.3 |
| VCG+VSG+TGKs | 93.3 | 82.0 |
| VCG+VSG+CGKs | **97.0** | **90.7** |

Table 1. The overall accuracy of the six approaches on the KTH and UCF sport datasets.

| | KTH (%) | UCF (%) |
|---|---|---|
| Yeffet *et al.* [24] | 90.1 | 79.2 |
| Wang *et al.* [22] | 92.1 | 85.6 |
| Kovashka *et al.* [11] | 94.5 | 87.3 |
| Le *et al.* [12] | 93.9 | 86.5 |
| Wang *et al.* [21] | 94.2 | 88.2 |
| Jiang *et al.* [10] | 95.8 | 88.0 |
| Wang *et al.* [23] | 93.3 | - |
| Celiktutan *et al.* [3] | 90.6 | - |
| Our 'VCG+VSG+CGKs' | **97.0** | **90.7** |

Table 2. The comparison of our approach with the state-of-the-art approaches on the KTH and UCF sport datasets.

be observed through analyzing the experimental results of the six different approaches.

First, the '0-CGK' approach achieves $94.7\%$ accuracy, which is $0.8\%$ higher than that of the 'BoW' approach. It indicates that without considering the spatio-temporal relationships among local features, our 0-order CGK still can achieve a relatively considerable performance on this dataset.

Secondly, the 'VCG+CGKs' and 'VSG+CGKs' approaches, reaching the accuracies of $95.5\%$ and $96.3\%$, are $0.8\%$ and $1.6\%$ higher than the '0-CGK' respectively. It demonstrates that both VCG and VSG do preserve spatio-temporal relationships among local features detected in the intra-frames and the inter-frames effectively. And these preserved spatio-temporal relationships, can improve the performance of human action recognition.

Furthermore, the 'VCG+VSG+CGKs' approach outperforms both 'VCG+CGKs' and 'VSG+CGKs' approaches and achieves performance of $97.0\%$. We can infer that the VCG and VSG are complementary to each other and combination of both graphs together will lead a sufficiently informative and discriminative representation for human actions.

Finally, comparing 'VCG+VSG+CGKs' with 'VCG+VSG+TGKs', we can see that the former is $3.7\%$ higher than the latter on the accuracy. It shows that the proposed CGKs is superior to the TGKs in measuring similarity between graphs. The context information utilized in PWG matching can improve the performance of CGKs in similarity measurement between graphs.

### 5.2. Experiments on the UCF Sports Dataset

The UCF sports dataset contains 150 broadcast sports videos of ten different types of actions. The collection of the dataset represents a natural pool of actions featured in a wide range of scenes and viewpoints, so the videos exhibit great intra-class variation. In the experiments, we take the leave-one-out cross validation, namely cycling each example as a test video one at a time. We also test all six approaches on this dataset. For the 'BoW' approach, the number of visual words is set to 800 based on the cross validation. The performances of different approaches on each class are shown in Figure 4 and the average accuracies are presented in Table 1.

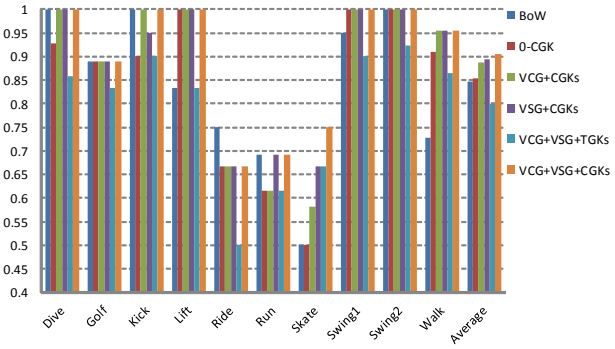


Figure 4. The performance of six approaches on the UCF sports dataset.

Table 1 and Figure 4, show that the 'VCG +VSG+CGKs' approach outperforms the other five approaches, reaching $90.7\%$ on the UCF sports dataset. Through analyzing the experimental results of different approaches, we draw the conclusions similar to that on the KTH dataset, which demonstrate the effectiveness of our approach on the realistic and complicated action dataset.

In addition, we also compare the performances of our approach with other state-of-the-art approaches on both the KTH and UCF sports datasets. The experimental results are shown in Table 2. It can be observed that our approach outperforms the listed approaches on both datasets.

### 5.3. Experiments on the UCF Films Dataset

The UCF films dataset provides a representative pool of natural samples of action classes, including kissing and slapping. There are 92 video sequences of kissing and 112 sequences of slapping. The video sequences are extracted from classic movies and appear in a wide range of scenes, viewpoints. We proceed a leave-one-out cross validation fashion in the experiments. The comparison between our 'VCG+VSG+CGKs' approach and other previous ap-

| | Kiss(%) | Slap(%) | Average(%) |
|---|---|---|---|
| Rodrigues *al et.* [15] | 66.4 | 67.2 | 66.8 |
| Yeffet *al et.*[24] | 77.3 | 84.2 | 80.7 |
| Our Approach | 97.6 | 94.4 | **96.0** |

Table 3. The comparison of our approach with the state-of-the-art approaches on the UCF films dataset.

proaches is shown in Table 3. It can be observed that our approach achieves a higher performance.

## 6. Conclusion

In this paper, we have proposed a new graph based approach for human action recognition. First we have constructed two complementary video graphs VCG and VSG to represent an action, capturing the spatio-temporal relationships among local features. Then a family of CGKs has been proposed for the similarity measurement between graphs. We have decomposed graphs into PWGs and different order CGKs have been computed based on the PWG matching. We have taken advantage of the contexts of PWGs to improve the performance of PWG matching. Finally, a GMKL with $l1$2-norm regularization has been applied to combine CGKs together and classify human actions. Experiments on several datasets have demonstrated that our two-graph model which preserves the spatial and temporal relationships among local features can improve the performance of action recognition and the proposed CGKs provide an efficient and multi-order similarity measurement on attributed graphs.

## 7. Acknowledgement

This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), the Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081) and NSFC (Grant No. 61100099, 61303086).

## References

[1] K. M. Borgwardt, C. S. Ong, S. Schönauer, S. Vishwanathan, A. J. Smola, and H.-P. Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21:47–56, 2005.

[2] E. Z. Borzeshi, M. Piccardi, and R. Xu. A discriminative prototype selection approach for graph embedding in human action recognition. In *ICCV*, pages 1295–1301, 2011.

[3] O. Çeliktutan, C. Wolf, B. Sankur, and E. Lombardi. Real-time exact graph matching with application in human action recognition. In *HBU*, pages 17–28. 2012.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VSPETS*, pages 65–72, 2005.

[5] T. Gärtner, P. Flach, and S. Wrobel. On graph kernels: Hardness results and efficient alternatives. In *LTKM*, pages 129–143. 2003.

[6] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury. A string of feature graphs model for recognition of complex activities in natural videos. In *ICCV*, pages 2595–2602, 2011.

[7] B. Gauzere, L. Brun, D. Villemin, and M. Brun. Graph kernels based on relevant patterns and cycle information for chemoinformatics. In *ICPR*, pages 1775–1778, 2012.

[8] Z. Harchaoui and F. Bach. Image classification with segmentation graph kernels. In *CVPR*, pages 1–8, 2007.

[9] W. Imrich, S. Klavžar, and B. Gorenec. *Product graphs: Structure and recognition*. Wiley New York, 2000.

[10] Z. Jiang, Z. Lin, and L. S. Davis. Recognizing human actions by learning and matching shape-motion prototype trees. *PAMI*, 34(3):533–547, 2012.

[11] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, pages 2046–2053, 2010.

[12] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, pages 3361–3368, 2011.

[13] M. Parsana, S. Bhattacharya, C. Bhattacharya, and K. Ramakrishnan. Kernels on attributed pointsets with applications. In *NIPS*, pages 1129–1136, 2007.

[14] K. Raja, I. Laptev, P. Pérez, and L. Oisel. Joint pose estimation and action recognition in image graphs. In *ICIP*, pages 25–28, 2011.

[15] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, pages 1–8, 2008.

[16] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, pages 32–36, 2004.

[17] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *ICMM*, pages 357–360, 2007.

[18] A. P. Ta, C. Wolf, G. Lavoue, and A. Baskurt. Recognizing and localizing individual activities through graph matching. In *AVSS*, pages 196–203, 2010.

[19] M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *ICML*, pages 1065–1072, 2009.

[20] S. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *JMLR*, 99:1201–1242, 2010.

[21] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, 2011.

[22] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[23] L. Wang, Y. Qiao, and X. Tang. Motionlets: Mid-level 3d parts for human motion recognition. In *CVPR*, 2013.

[24] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, pages 492–497, 2009.