Spatio-Temporal Context Kernel for Activity Recognition

Fei YUAN*, Hichem Sahbi†, Veronique Prinet*
*National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, P.R.C
Email: {fyuan,vprinet}@nlpr.ia.ac.cn
†LTCI, CNRS, TELECOM ParisTech, Paris, France
Email: sahbi@telecom-paristech.fr

***Abstract*—Local space-time features and bag-of-feature (BOF) representation are often used for action recognition in previous approaches. For complicated human activities, however, the limitation of these approaches blows up because of the local properties of features and the lack of context. This paper addresses the problem by exploiting the spatio-temporal context information between features. We first define a spatio-temporal context, which combines the scale invariant spatio-temporal neighberhood of local features with the spatio-temporal relationships between them. Then, we introduce a spatio-temporal context kernel (STCK), which not only takes into account the local properties of features but also considers their spatial and temporal context information. STCK has a promising generalization property and can be plugged into SVMs for activities recognition. The experimental results on challenging activity datasets show that, compared to context-free model, the spatio-temporal context kernel improves the recognition performance.**

I. Introduction

Human activity recognition is one of the most challenging problems in computer vision. By "activity", we refer to a higher-level combination of primitive actions with certain spatial and temporal relationships, *e.g.*, hand shaking, hugging, eating food with a silverware, *etc*. Compared with simple actions, such as walking and drinking, activity recognition is more complicated. Besides general difficulties of action recognition, such as camera motion, illumination changes, occlusion, low intra-similarity and large inter-variability, *etc*, the challenge of activity recognition stems from its structured property: the complicated spatio-temporal interactions between a set of body parts or multiple persons.

Bag-of-feature (BOF) representation have been widely employed for action recognition, both for local spatio-temporal features [1]–[3] and trajectory-based features [4]–[7], since it gives a sparse and simplified representation of actions and can be effectively integrated into a machine learning framework. Impressive results have been reported in both synthetic and realistic scenarios, see [1]–[4], [8]–[10]. However, the limitation of bag-of-feature (BOF) representation blows up when they are used to represent complicated activities with long-range motions or multiple interactive body parts. Because the bag-of-features representation makes an important assumption that all the features are independent and thus the geometrical and temporal relationships among features are discarded.

Context has been considered as an important cue for activity recognition. [11] proposed to learn the shapes of neighborhoods of the space-time features which are discriminative for a given action category, and recursively mapped the descriptors of the variable-sized neighborhoods to higher-level vocabularies. The process produced a hierarchy of space-time configurations. In [12], the objects and human body parts were considered as mutual context, and then a random field model was proposed to encode the mutual context of objects and human poses in "human-object interaction" activities. They casted the model learning task as a structure learning problem, by which the structural connectivity between objects, overall human poses, and different body parts were estimated.

In this paper, we introduce a Spatio-Temporal Context Kernel (STCK) to exploit the structural and dynamic property of activity features for activity recognition . We argue that an activity is composed of a set of activity features, which depend on each other both in the spatial domain and in the temporal domain. In STCK, one activity feature is considered to be related to all other features in its neighbourhood, thus the similarity of two activities is computed by reling on both the local properties of features and the spatio-temporal context of features. Moreover, STCK has a promising generalization property and can be plugged into SVMs for activities recognition.

We first present a spatio-temporal context, which combines the scale invariant spatio-temporal neighberhood of local features with the spatio-temporal relationships between them. The spatio-temporal context devides the spatio-temporal neighberhood of a local feature into a set of sub-contexts. Different sub-contexts describe different context information, which makes it possible to consider the context in a precise way. We then introduce a Spatio-Temporal Context Kernel (STCK). In STCK, the matching of two features in two videos is related to the matching of features around its context zone, which is achieved by a context term. By the context term, a high value of two features in two videos implies high kernel values in their context zone.

The rest of this paper is organized as follows. Sec. II gives a detailed description of our spatio-temporal context. In Sec. III, we present the Spatio-Temporal Context Kernel (STCK) for activity representation. We illustrate and interpret the experimental results in Sec. IV, and finally conclude the paper in Sec. V.

978-1-4577-0121-4/11/$26.00 ©2011 IEEE

II. Spatio-Temporal context

The bag-of-features activity representation makes an important assumption that all the activity features in a video clip are independent. Observe that this assumption has been also widely adopted, and few work considers the relationships between the activity features. However, one basic fact ignored in activity recognition is that an activity can be composed of several activity features, such as the activity *shaking hand* consisting of two shaking hands and two moving arms, which depend on each other not only in the spatial domain but also in the temporal domain. Thus, in this section, we introduce the spatio-temporal context of activity features for taking into account the relationships between them.

A. Activity Components

In this work, we use a mid-level feature proposed in [13], named activity-components, as the basic features to represent activity. An activity-component is a connected spatio-temporal part having consistent spatial structure and consistent motion in temporal domain. They are extracted by clustering similar point-trajectories according to their appearance and motion attributes. See [13] for more details about this feature.

Activity-components were reported being more discriminative than local space-time interest features [13], and the spaseness of activity-components makes it effective and efficient to consider the context information between them.

B. Spatio-temporal neighborhood

Let note $\mathcal{V}$ an activity represented by a collection of activity-components as $\mathcal{V} = \{c_n\}_{n=1}^{N}$. In our case, each activity-component c is described by a vector $\langle f(c),\ \ell(c),\ s(c)\rangle$, where $f(c)$ stands for the feature of the activity-component as detailed in Section II-A, $\ell(c) = \big(\overline{x}(c), \overline{y}(c), \overline{t}(c)\big)$ corresponds to the 3-dimensional centroid location of c, which is computed from all the key-points in the component. Let $s(c) = \big(s_{xy}(c), s_t(c)\big)$ be the spatial scale and temporal scale of c, where $s_{xy}(c)$ and $s_t(c)$ are the spatial and temporal average radius of the component, respectively.

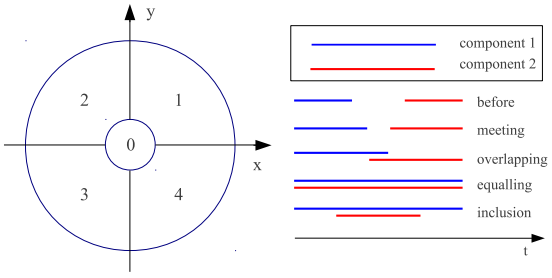
Given a component c with centroid location $\ell(c) = \big(\overline{x}(c),\ \overline{y}(c),\ \overline{t}(c)\big)$, note $\ell_{xy}(c) := \big(\overline{x}(c),\ \overline{y}(c)\big)$, $\ell_t := \overline{t}(c)$ and define the spatial and temporal distance between two components c and c' as

$$d_{xy}(c, c') := \|\ell_{xy}(c) - \ell_{xy}(c')\|_2,$$
$$d_t(c, c') := \|\ell_t(c) - \ell_t(c')\|_2,$$

where $\|\cdot\|_2$ is the L_2-norm. Thus, the spatio-temporal neighborhood of c is defined in a way proportional to the spatial and temporal scale of the component

$$\mathcal{N}(c) := \Big\{c' : c' \in \mathcal{V},\ d_{xy}(c, c') < \alpha_{xy} \cdot s_{xy}(c), \\ d_t(c, c') < \alpha_t \cdot s_t(c)\Big\}, \quad (1)$$

where α_{xy} and α_t are the neighborhood factors. Obviously, $\mathcal{N}(c)$ is a circular cylinder with a radius $\alpha_{xy} \cdot s_{xy}$ and a half length $\alpha_t \cdot s_t(c)$.



(a) Spatial relationships (b) Temporal relationships

Fig. 1. Illustration of spatio-temporal relationships. In (a), the spatial relationships are quantized into 5 states. In (b), the locations of lines stand for the temporal locations of components, and the lengths of lines stand for the temporal ranges components cover.

C. Spatio-temporal relationships

We describe the spatial and temporal relationships between components respectively and quantize them into discrete states. (1) In order to consider the spatial relationships between two components c and c', we first compute the spatial distance $d_{xy}(c, c')$ and then quantize the relationship into 5 different states as demonstrated in Figure 1(a). More precisely, if $d_{xy}(c, c')$ is smaller than a given value λ, then relationship is quantized as state 0; otherwise the angle of the spatial locations of two components (*i.e.* the direction from $\ell_{xy}(c)$ to $\ell_{xy}(c')$) is calculated and is quantized into the other four states. (2) For temporal relationships, similar to [14], we define 5 types of relationships based on the temporal locations $\ell_t(c), \ell_t(c')$ and the temporal scales $s_t(c), s_t(c')$ of the components c and c' respectively: *before*, *meeting*, *overlapping*, *equaling*, *inclusion*. See Figure 1(b) for a graphical illustration. Thus, there are $5 \times 5 = 25$ spatio-temporal relationships in total.

These spatio-temporal relationships reflect the spatial structure arrangement and the temporal dynamic information of human activities, which can be used as an important cue to help understanding human motions and recognizing activities.

D. Context with spatio-temporal relationships

According to the spatio-temporal relationships mentioned above, we define the context of an activity-component c as

$$\mathcal{N}_r(c) := \Big\{c' : c' \in \mathcal{N}(c),\ c' \curlywedge c = r\Big\},\ r \in \{0, 1, \ldots, 24\}, \quad (2)$$

where r is the label of spatio-temporal relationship, $c' \curlywedge c$ stands for the quantized spatio-temporal relationship, and $\mathcal{N}(c)$ is the neighborhood of the component c.

III. Design of spatio-temporal context-dependent kernels

In order to model the spatio-temporal context of activity-components, we rely on the context-dependent kernel, which was first proposed by Sahbi *et al.* [15] for capturing geometric structure information of objects. In this section, we extend the context-dependent kernel into spatio-temporal domain for exploiting the structural and dynamic property of activities.

Let $\mathcal{C}$ be the union of possible components over the activities $\{\mathcal{V}_1, ..., \mathcal{V}_N\}$. We consider $K : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}^+$ as a kernel which provides a similarity measure between any two activities $\mathcal{V}_p = \{c_i^p\}_{i=1}^m$ and $\mathcal{V}_q = \{c_j^q\}_{j=1}^n$ as

$$K(\mathcal{V}_p,\ \mathcal{V}_q) = \sum_{i=1}^{m}\sum_{j=1}^{n} k(c_i^p, c_j^q),$$

where $k(c_i^p, c_j^q)$ is a kernel measuring the similarity between two components c_i^p and c_j^q. Let $\{P_r(c_i^p, c_j^q) = g_r(c_i^p, c_j^q)\}_{c_i^p, c_j^q}$ be the elements of intrinsic adjacency matrices with a spatio-temporal relationship label r, where $g_r(c_i^p, c_j^q)$ is a decreasing function of any (pseudo) distance involving components c_i^p and c_j^q. Let $D(c_i^p, c_j^q) = d_f(c_i^p, c_j^q)$, where $d_f(c_i^p, c_j^q)$ is a dissimilarity metric between components c_i^p and c_j^q in feature space $\{f\}$. According to [15], the kernel K on $\mathcal{C} \times \mathcal{C}$ could be defined by minimizing

$$\min_{K\geq 0, \|K\|_1=1} \mathrm{Tr}\big(KD^T\big) + \beta \mathrm{Tr}\big(K \log K^T\big) - \alpha \sum_{r=0}^{24} \mathrm{Tr}\big(KP_rK^TP_r^T\big), \qquad (3)$$

where D^T is the transposed matrix of D, $\alpha, \beta \geq 0$ are parameters and the operations "log"(natural) and "$\geq$" are applied individually to every entry of the matrix, $\|\cdot\|_1$ is the "entry wise" L_1-norm and Tr denotes matrix trace.

The first term in the above constrained minimization problem is a fidelity term which measures the quality of how the features of components match. The second term is a regularization term which keeps the probability distribution $\{k(c_i^p, c_j^q)\}$ flat without any priori knowledge about the aligned components. The third term is a term capturing the spatio-temporal structures, where a high value of $\{k(c_i^p, c_j^q)\}$ should imply high kernel values in the neighborhoods $\mathcal{N}_r(c_i^p)$ and $\mathcal{N}_r(c_j^q)$.

In practice, we use Euclidean distance to compute the dissimilarity of features

$$d_f(c_i^p, c_j^q) = ||f(c_i^p) - f(c_j^q)||_2,$$

and take the distance function

$$g_r(c_i^p, c_j^q) = \sum_{c_k^p \in \mathcal{N}_r(c_i^p),\ c_l^q \in \mathcal{N}_r(c_j^q)} g_r(c_i^p, c_k^p) \cdot g_r(c_l^q, c_j^q),$$

where $g_r(c_i^p, c_k^p) = \exp\{-\frac{1}{\sigma_d} d_\ell(c_i^p, c_k^p)\}$ with $d_\ell(c_i^p, c_k^p) = \|\vec{\omega} \cdot \big(\ell(c_i^p) - \ell(c_k^p)\big)\|_2$ as a weighted spatial and temporal distance between two neighbors, $\vec{\omega} = (\omega_x, \omega_y, \omega_t)$.

a) Kernel solution: The optimization problem in Equation (3) admits a solution K, which is the limit of the context-dependent kernels

$$K^{(\eta)} = \frac{G(K^{(\eta-1)})}{\sum_{i,j} G(K^{(\eta-1)})},$$

with

$$G(K^{(\eta)}) = \exp\Big\{ -\frac{D}{\beta} + \frac{\alpha}{\beta} \sum_{r=0}^{24} \big(P_r K^{(\eta-1)} P_r^T + P_r^T K^{(\eta-1)} P_r\big)\Big\},$$

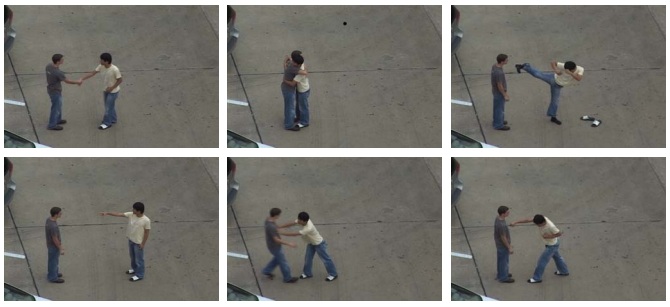
$$K^{(0)} = \frac{\exp(-\frac{D}{\beta})}{\sum_{i,j} \exp(-\frac{D}{\beta})}.$$

Refer to [15] for detailed proof of this solution and its convergence to a positive definite fixed point. In practice, the kernel usually stops after 3 iterations, *i.e.* $\eta = 2$.

So far, two activities $\mathcal{V}_p$ and $\mathcal{V}_q$ can be compared within spatio-temporal context by designed spatio-temporal kernel $K(\mathcal{V}_p,\ \mathcal{V}_q)$. It is worth noticing that this kernel can be plugged into classifiers, such as SVMs, which benefits both the context-dependent kernel and the well established generalization power.

IV. Experimental results

This section experimentally evaluates the proposed method. we demonstrate that taking into account the spatio-temporal context of activity-components through STCK can largely improve the performance of activity recognition.



(a) Snapshots of video sequences in the UT-Interaction dataset [16], containing 6 classes of activities and 20 videos in each class.

(b) Snapshots of video sequences in the Rochester Activities Dataset [5], containing 10 classes of activities and 15 videos in each class.

Fig. 2. Snapshot examples of video sequences in two activity datasets: the UT-Interaction dataset [16] and the Rochester Activities Dataset [5].

A. Experimental Setups

All the comparisons are implemented on two recent activities datasets: the UT-Interaction dataset [16] containing complex interactions and the Rochester Activities Dataset [5] containing complicated daily activities, Both of which are reported to be challenging on activity recognition.

b) UT-Interaction dataset: It contains 6 classes of human-human interactions: *shake-hands, point, hug, push, kick* and *punch*. Each class contains 20 video sequences which are divided into two different sets: (1) the first 10 video sequences, Set 1, are taken on a parking lot with slightly different zoom rate, and the backgrounds of the video are mostly static with little camera jitter; (2) the other 10 video sequences, Set 2, are taken on a lawn in a windy day. The backgrounds of the video are moving slightly, *e.g.* tree moves, and the videos contain more camera jitters. Observe that the background, scale, and illumination of the videos in each set are different. Figure 2(a) shows one snapshot for each class.

c) Rochester Activities Dataset: This dataset contains 10 classes of daily living activities: *answering a phone, chopping a banana, dialing a phone, drinking water, eating a banana, eating snack chips, looking up a phone number in a telephone book, peeling a banana, eating food with silverware* and *writing on a white board*. Each of these activities contain 15 different sample videos, which were performed three times by five different people of different shapes, sizes, genders, and ethnicities. It has been reported that only using motion information is not sufficient for distinguishing these activities and some other information, such as appearance descriptions, should be taken into account [5]. Figure 2(b) shows one snapshot for each class on the dataset.

In order to evaluate and compare the performances, we duplicate experimental setting in [5] and [16]. Specifically, for Rochester Activities Dataset, 12 video sequences performed by four out of the five subjects are used as training set, and the left 3 video sequences are used for testing. The experiments are repeated five times; For UT-Interaction dataset, a 10-fold leave-one-out cross validation (each times, 9 samples are used for training and the left one is used for testing) is implemented on each set. Different methods are compared by using the average recognition rates.

B. Comparative Evaluations

To evaluate the effect of the proposed method, we model and classify activities with 3 representations:

- *Bag-of-Component + SVM*: we first generate a codebook of activity-components using k-means, then represent activity with bag-of-features model, finally train the model of each class by SVMs with radial basis function kernel. The size of codebook is 600 for both datasets.
- *Component + STCK + SVM*: The model of each class is trained by SVMs with the STCK. When computing the STCK, the following parameters are used: $\alpha_{xy} = 4$, $\alpha_t = 6$, $\alpha = 20$ and $\beta = 0.5$.
- *Component + context-free kernel + SVM*: The modeling and classification is the same as in *Component + STCK + SVM*. However, in this case, $\alpha = 0$ is used, which implies that no context is taken into account when computing the kernels.

The above parameters are chosen in order to achieve the best performance on the databases.

The STCK is compared with 2 context-free models: the bag-of-components model and the context-free kernel. As mentioned before, the context-free kernel is the special case of STCK, that is $\alpha = 0$. Equal error rate (EER) is used as a measure to evaluate the performances, which equally weights the positive and the negative errors. The smaller the EER, the better the performance is.

Figure 3 shows the EER of activities in UT-Interaction dataset. We can see that *Component + STCK + SVM* outperforms *Bag-of-Component + SVM* and *Component + context-free kernel + SVM*, both on Set 1 and Set 2. Specifically, on Set 1, the average EER of *Bag-of-Component + SVM* and *Component + context-free kernel + SVM* are 24.2% and 33.3% respectively, but for the proposed *Component + STCK + SVM*, we get the EER of 19.2%. The proposed spatio-temporal context-dependent model *Component + STCK + SVM* also improves two context-free model in Set 2, which outperforms *Bag-of-Component + SVM* (32.5%) by 6.7% and *Component + context-free kernel + SVM* (38.3%) by 12.5% respectively.

Figure 4 shows the EER of activities in Rochester Activities Dataset. We can see that the EER of *Component + STCK + SVM* is lower than those of *Bag-of-Component + SVM* and *Component + context-free kernel + SVM* in 8 activity classes and is as the same as those in lookupInPB and writeBoard classes. For the acerage EER, the proposed *Component + STCK + SVM* outperforms *Bag-of-Component + SVM* (17.3%) by 8% and *Component + context-free kernel + SVM* (25%) by 14.7% respectively.

The above experiments show that the performances of STCK are consistently better than those of the context-free kernel and the bag-of-components model on the two datasets.

V. Conclusion

We have presented a Spatio-Temporal Context Kernel (STCK) for activity recognition, which not only takes into account the local properties of features but also considers their spatial and temporal context information. We demonstrated that, when using proper context, STCK improves the matching between features in different video. Experiments on challenging activity datasets have confirmed that our proposed STCK outperforms the popular bag-of-features representation and the context-free kernel.

References

[1] I. Laptev, , and T. Lindeberg, "On space-time interest points," in *Proc. Int. Conference on Computer Vision (ICCV)*, 2003.

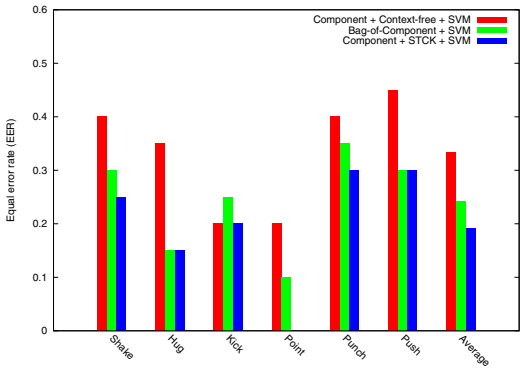
[2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. Int. Conference on VS-PETS*, 2005.

[3] G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. European Conference on Computer Vision (ECCV)*, 2008.

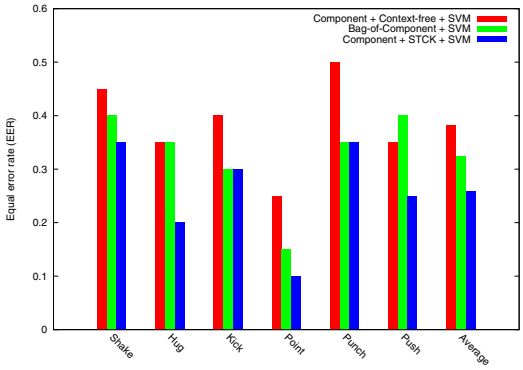
[4] P. Siva and T. Xiang, "Action detection in crowd," in *Proc. British Machine Vision Conference (BMVC)*, 2010.

[5] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Proc. Int. Conference on Computer Vision (ICCV)*, 2009.

[6] P. Turaga and R. Chellappa, "Locally time-invariant models of human activities using trajectories on the grassmannian," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.



(a) UT-Interaction dataset set 1



(b) UT-Interaction dataset set 2

Fig. 3. Equal error rate (EER) on UT-Interaction dataset set 1 and set 2.

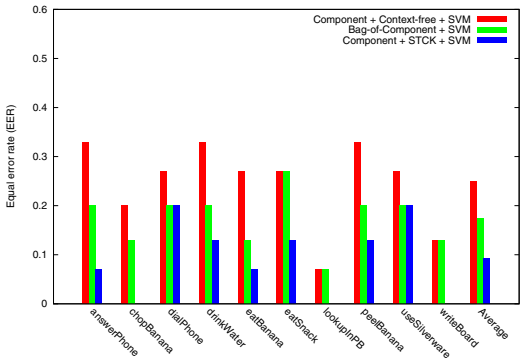


Fig. 4. Equal error rate (EER) on on Rochester Activities Dataset.

[7] H. Wang, A. Kläser, C. Schmid, and L. Cheng Lin, "Action Recognition by Dense Trajectories," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[8] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[9] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[10] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[11] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[12] B. Yao and L. Fei-Fei, "Modeling mutual context of object and human pose in human-object interaction activities," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[13] F. Yuan, J. Yuan, and V. Prinet, "Middle-level representation for human activities recognition: the role of spatio-temporal relationships," in *ECCV'10 Workshop on Human Motion: Understanding, Modeling, Capture and Animation*, 2010.

[14] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *Proc. Int. Conference on Computer Vision (ICCV)*, 2009.

[15] H. Sahbi, J.-Y. Audibert, J. Rabarisoa, and R. Keriven, "Context-dependent kernel design for object matching and recognition," in *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[16] M. S. Ryoo and J. K. Aggarwal, "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)," http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.