Action Recognition with Multiscale Spatio-Temporal Contexts

Jiang Wang , Zhuoyuan Chen and Ying Wu
EECS Department, Northwestern University
2145 Sheridan Road, Evanston, IL 60208
{jwa368,zch318,yingwu}@eecs.northwestern.edu

Abstract

The popular bag of words approach for action recognition is based on the classifying quantized local features density. This approach focuses excessively on the local features but discards all information about the interactions among them. Local features themselves may not be discriminative enough, but combined with their contexts, they can be very useful for the recognition of some actions. In this paper, we present a novel representation that captures contextual interactions between interest points, based on the density of all features observed in each interest point's mutliscale spatio-temporal contextual domain. We demonstrate that augmenting local features with our contextual feature significantly improves the recognition performance.

1. Introduction

Action recognition and categorization in video have received great attention because of its difficulty and potential applications. It can be used in human-computer interface, video surveillance, video indexing, and many other areas. Recently, various approaches have been proposed and many progresses have been achieved. However, it still remains a challenging problem due to several fundamental difficulties.

The first difficulty arises from the tremendous intra-pattern variations in human actions. The same type of action can have huge differences in their visual appearance, variations in performing speed, clothing, and viewpoints.

The spatio-temporal nature of the data also adds to the complexity. Modeling spatio-temporal dependencies for 3D video data poses a big challenge. To simplify the task, researchers usually resort to schemes that assume conditional independence across spatial and temporal domains such as the bag of words method, which summarizes a video by the histogram of feature occurrences. Recently, many bag of words methods using local spatio-temporal descriptors have been proposed for action recognition tasks [20, 12, 16, 5, 19, 9]. Such descriptors are useful in representing the local appearance of the 3D salient points. They are robust to viewpoint and scale changes, simple to implement, and have achieved promising results.

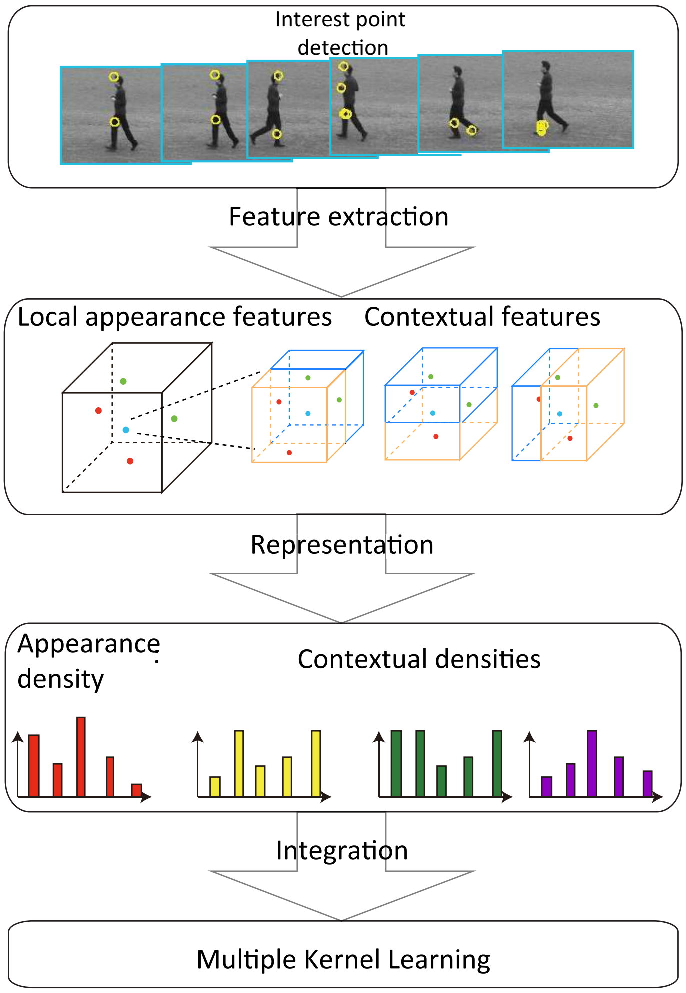

Figure 1. A schematic framework of spatio-temporal contextual feature

However, the bag of words methods have their limitations. One important limitation is that they largely ignore the spatio-temporal relationship among the local appearance features, such as temporal order of the action, spatial arrangement of objects, and motion trajectories. For example, in Fig. 2, the actions in the left and right video volumes are considered the same in the bag of words method. However, the semantic meaning of these two actions may be different due to the spatio-temporal configuration differences. In order to solve this problem, we introduce a novel contextual feature to incorporate the spatio-temporal dependencies into the bag of words model.

The term *context* is widely used in object recognition area. Contextual information can be very important for action recognition, because an action is characterized by the interaction between human parts or between human and objects. In the action categorization field, the context usually refers to the information about typical configurations of objects in a scene. It can be either global or local. The global context usually exploits scene configuration as an extra source of global information across categories [26]. Local context captures the local arrangement of the objects or regions [36, 21]. Both of those contextual features have proved to be beneficial to action recognition tasks, but most of the existing contextual features tend to be complicated.

We consider the problem of designing simple but discriminative local contextual descriptors in this paper. Traditional contextual features either use object detection [22] or segmentation [21] as preprocessing, or require complex learning methods such as AdaBoost [9] or discriminative configuration mining [12]. Although these approaches have achieved good recognition results, they have two shortcomings. First, these processing procedures are time-consuming. Second, because the object recognition and machine leaning procedures are error-prone, especially when the data is noisy, using these procedures may add noise to the results.

This paper presents a new approach to incorporating contexts into action recognition. In the proposed contextual model, there are some feature classes, each of which is associated with an individual context. One individual context for one pixel is represented by the posterior density of this particular feature class at this pixel location, based on the density of all features observed in its multiscale spatio-temporal contextual domains. Its total context is the collection of these individual ones. As this density is able to reflect the change in the action-object interaction, contextual features exhibit more discriminative power compared to traditional local appearance features.

In our approach, we extract spatio-temporal extrema points as interest points. Multiple channels of contextual features are generated for each interest point. For computational efficiency and performance consideration, multiple kernel learning is used to select the best combination of channels in a multi-channel SVM classification. An illustration of the framework of the proposed method is shown in Fig. 1.

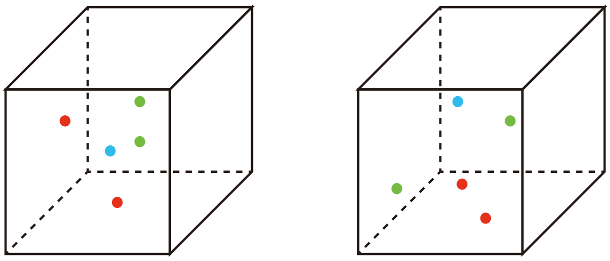

Figure 2. Two spatio-temporal video volumes with interest points in them. The color of the points represents the feature classes .A simple bag of words approach will consider both volumes to be identical as it does not consider the spatio-temporal configuration. But the two volumes may express different actions.

The novelty of this paper includes the following aspects: (1) It augments local appearance features with contextual method as features that represent the relationship among the 3D salient points. (2) It utilizes an efficient classification method to integrate contextual features with local features.

We apply our approach to learn human action categories on the KTH [32] and the ADL [28] datasets, and show that the proposed method which includes contextual features is able to outperform some state of the art methods.

After a brief review of the related work in Sec.2, the description of the contextual feature is given in Sec. 3. We introduce our classification framework that integrates local appearance feature and contextual features.. The experimental results are reported in Sec. 5. Sec. 6 concludes this paper.

2. Related Work

Action recognition is an important topic in computer vision field. Researchers have proposed many different kinds of approaches to solve this problem. One type of approaches uses motion trajectories to represent actions [34, 31, 18, 28], and requires target or feature tracking. Another type of approaches uses sequences of silhouettes or body contours to model actions [23, 30, 24, 23], and this type of methods require background subtraction. Recently, action categorization method that uses local spatio-temporal features has drawn a lot of attention, because of its robustness to viewpoint and scale change, as well as superior performance [20, 13, 16, 5, 19, 38].

The context information can be used to help object and action recognition problem. [11] gives a comprehensive review of the contextual features and machine learning models that integrate contextual information into object recognition frameworks. Previous contextual features in action recognition include the scene information [25, 22], spatio-temporal relationship between trajectories

[34], neighborhood-based feature [17], spatial visual feature arrangement [36] and object-level interaction characterized by graphical models [10, 37, 14].

Different from those approaches, our method represents the contextual information as spatio-temporal statistics in the 3D neighborhood of each interest point. Other recent work also used feature-centered approach [9, 13], but employed different representations. [13] uses co-occurrence transaction to describe context feature, which may suffer from the loss of information. [9] uses AdaBoost to learn a classifier in each interest point's local region, and uses classification score as contextual feature. [27] represents the relationship between local features as the distribution of quantized location difference between each pair of interest points.

Some work in object recognition is related to our proposed work as well. In [1], shape context at the represent point captures the distribution of the remaining points relative to it, thus offering a globally discriminative characterization. In [35], local feature statistics is used to compute more accurate optical flow. In [39], a self-supervised clustering algorithm is proposed with the help of local contextual information.

3. Contextual Feature Model

We propose to capture local contextual information by spatio-temporal statistics. Given a video sequence, we first extract spatio-temporal interest points $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_P\}$, then accumulate the local features around each interest point's spatio-temporal contextual domain, and obtain different sets of contextual features:$\boldsymbol{H} = \{H_1(\boldsymbol{x}), \ldots H_M(\boldsymbol{x})\}$.

In the following, we give a detailed description of our approach. We start with the spatio-temporal interest points extraction.

3.1. Spatio-temporal Interest Points

Spatio-temporal features have shown good performance for action recognition[20]. They provide a compact representation for video, and achieve robustness against intra-class variations. To detect an interest point, we use the Harris3D corner detector [20], which is an extension of the 2D Harris corner. Matrix $\boldsymbol{U}$ is defined as:

$$\boldsymbol{U} = \begin{bmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{bmatrix} \tag{1}$$

where L_x, L_y and L_t are the gradients of Gaussian-smoothed video in horizontal, vertical and temporal direction. The Harris3D corner detector finds points whose $\boldsymbol{U}$ has large eigenvalues. Those points are generally extrema points in the spatio-temporal space.

For each interest point, we extract the histogram-of-oriented-gradients (HOG) [7] and histogram-of-optical-flow (HOF) features [8], which characterize the local appearance (including visual appearance and motion appearance) of the interest points. These features are called local features in this paper.

The local features are clustered into N classes by k-means algorithm. The N centers of these clusters are called visual words. Each local feature can be mapped into a visual word.

3.2. Contextual Features

Consider a spatiotemporal point that is located at $\boldsymbol{x} = [u, v, t]$, where u, v and t are its horizontal, vertical and temporal coordinates. The point is associated with a local feature vector $\boldsymbol{f}(\boldsymbol{x}) \in \mathcal{R}$ in a feature space. The feature space has been quantized to a finite set of N discrete feature classes $\{\omega_1, \ldots, \omega_N\}$, each of which is represented by its local appearance visual word.

A point is not isolated but surrounded by its spatial and temporal context. In our approach, P multiscale channels of contextual features are computed. Different channels of contextual features are computed within cuboids that are of different sizes and shapes. We aim to capture various types of interactions by using multiple channels of contextual features. For example, contextual domains with short spatial support and long temporal support can capture temporal evolution information.

For example, as shown in Fig. 3, we have three channels of contextual features. These contextual features have contextual domains with different shapes, and can capture different types of contextual information.

For each channel of contextual features, a regular grid is used to encode spatio-temporal information within the local neighborhood of an interest point. Each cuboid in this regular grid defines a contextual domain. The context for feature class ω_i in the jth contextual domain is defined as

$$C_{ij} = \{\boldsymbol{y} | f(\boldsymbol{y}) \in \omega_i, y \in \Omega_j(\boldsymbol{x})\} \tag{2}$$

where $\Omega_j(\boldsymbol{x})$ is jth spatio-temporal contextual domain of x with predefined size. This definition follows the definition in [35].

The total context in jth contextual domain $\Omega_j(\boldsymbol{x})$ is the union of all individual context inside it, *i.e.*

$$C_j = \bigcup_{i=1}^{N} C_{ij} \tag{3}$$

Because we are interested in the spatio-temporal configuration of the interest points. To represent the interaction around point $\boldsymbol{x}$, we use each feature class's posterior density at $\boldsymbol{x}$: $p(\omega_i|\boldsymbol{x})$. We have:

$$p(\omega_i|\boldsymbol{x}) = p(\boldsymbol{x}|\omega_i)p(\omega_i) \tag{4}$$

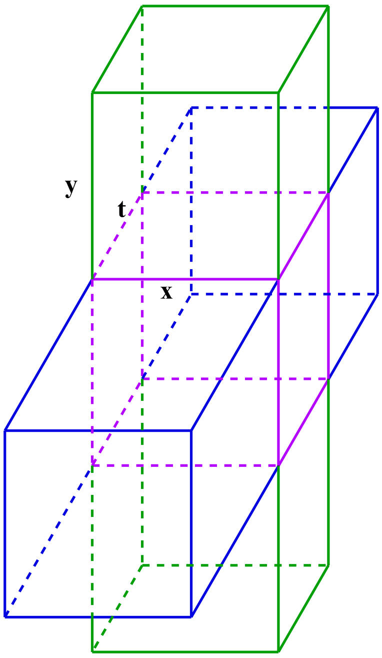

Figure 3. Three channels of contextual features. The purple contextual feature channel's contextual domains have x,y and t scales, while the blue and green contextual feature channel's contextual domain has larger support in t and y directions, respectively.

where the prior $p(\omega_i)$ can be estimated within a contextual domain, and $p(\boldsymbol{x}|\omega_i)$ can be computed using the kernel density estimation:

$$p(\boldsymbol{x}|\omega_i, C_j) = \sum_{\boldsymbol{x}_k \in C_{ij}} K(\boldsymbol{x}, \boldsymbol{x}_k) \tag{5}$$

where $K(.,.)$ is a 3D Gaussian kernel:

$$K(\boldsymbol{x}, \boldsymbol{x}_k) \propto \exp\left(-\left(\frac{(x-x_k)^2}{\delta_x^2} + \frac{(y-y_k)^2}{\delta_y^2} + \frac{(t-t_k)^2}{\delta_t^2}\right)\right) \tag{6}$$

where δ_x, δ_y and δ_t are independent spatial and temporal scale parameters.

For each contextual features channel of interest point $\boldsymbol{x}$, a regular grid is used to encode spatio-temporal information about the local neighborhood of an interest point. Specifically, each point has $M = W_x \times W_y \times W_t$ cuboids around it. Each cuboid corresponds to a spatio-temporal contextual domain Ω_j. The contextual feature contains M N-dimensional histogram vector $\{H_1(\boldsymbol{x}), \ldots H_M(\boldsymbol{x})\}$. Every histogram $H_j(\boldsymbol{x}) = \{b_{1j}, \cdots, b_{Nj}\}$ accumulates the probability of distance-weighted occupancies of each feature class ω_i within a given contextual domain, *i.e.*,

$$b_{ij}^0 = \sum_{\boldsymbol{x}_k \in C_{ij}} K(\boldsymbol{x}, \boldsymbol{x}_k) \tag{7}$$

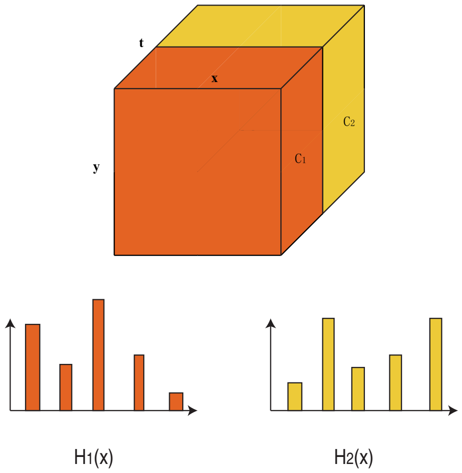

Figure 4. A simple contextual feature example with $W_x = 1, W_y = 1, W_t = 2$. The left and right histograms are the distance-weighted probability of occupancies of each feature class in the red and yellow regions respectively. The final contextual feature is the concatenation of these two histogram.

and

$$b_{ij} = b_{ij}^0 / \sum_{i=1}^{N} b_{ij}^0. \tag{8}$$

These M histograms encode the probabilities $p(\boldsymbol{x}|\omega_i)$ in all contextual domains. We concatenate M histograms to form one channel of our final contextual descriptor for a given point $\boldsymbol{x}$:$\{H_1(\boldsymbol{x}), \ldots H_M(\boldsymbol{x})\}$. With different selections of W_x, W_y, W_t, we can obtain multiple channels of contextual features. These contextual features characterize the relationship between the interest points.

For example, if $W_x = 1, W_y = 1, W_t = 2$ for a contextual feature channel, there are two contextual domains for an interest point $\boldsymbol{x}$ along temporal direction, as illustrated by red and yellow cuboids C_1 and C_2 in Fig.4. We accumulate the probabilities of distance-discounted occurrences for each feature class in C_1 and C_2, and obtain two histograms $H_1(\boldsymbol{x})$ and $H_2(\boldsymbol{x})$. This channel of contextual feature for $\boldsymbol{x}$ is $\{H_1(x), H_2(x)\}$.

In order to show the properties of our approach, we examine the behavior of our contextual features in two cases: First, we consider two actions: “catch a ball” and “throw a ball”. The contextual features for those two actions are different if $W_t > 1$, because the two actions have different temporal directions. Second, the “playing high bar” and “ walking” have similar local motion, but can be discriminated by contextual features with $W_y > 1$, as these actions have different spatial orientations. In conclusion, our contextual feature is descriptive of the spatio-temporal interactions of objects or features.

4. Integrating Contextual Features

4.1. Multiclass Multiple Kernel Learning

The channels of contextual features essentially share some information with each other. Training with all these features is computationally expensive. Thus, we use a linear combination of these features for efficiency. To optimize the combination of the features is a feature selection problem. Multiple kernel learning (MKL) [4] has proved to be a very good way of optimizing kernel weights while training a SVM. Since in our application, more than two classes are to be distinguished, the multiclass multiple kernel learning[40] is used.

A common approach to multiclass classification is the use of joint feature map $\Phi(\boldsymbol{x}, y)$ on data $\mathcal{X}$ and labels $\mathcal{Y}$ together with a linear output function

$$f_{\boldsymbol{w},\boldsymbol{b}}(\boldsymbol{x}, y) = \langle \boldsymbol{w}, \Phi(\boldsymbol{x}, y) \rangle + b_y, \tag{9}$$

parameterized with the hyperplane normal $\boldsymbol{w}$ and biases $\boldsymbol{b}$. The predicted class y for a point $\boldsymbol{x}$ is chosen to maximize the output.

$$\boldsymbol{x} \mapsto \arg\max_{y \in \mathcal{Y}} f_{\boldsymbol{w},\boldsymbol{b}}(x, y). \tag{10}$$

Multiclass-MKL considers a convex combination of p kernels, $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{p} \beta_k K_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Equivalently, we consider linear combinations of the corresponding output functions:

$$f_{\boldsymbol{w},\boldsymbol{b}}(\boldsymbol{x}, y) = \sum_{k=1}^{p} \beta_k \langle \boldsymbol{w}, \Phi(\boldsymbol{x}, y) \rangle + b_y. \tag{11}$$

We aim at choosing $\boldsymbol{w} = (\omega_w)_{k=1,\ldots,p}$ and $\boldsymbol{\beta}$ such that $f_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\beta}}(x_i, y_i) \geq f_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\beta}}(x_i, u)$ for all $u \in \mathcal{Y} - \{y_i\}$. The resulting optimization problem becomes:

$$\begin{aligned} &\min_{\beta,\boldsymbol{w},\boldsymbol{b},\xi} \frac{1}{2} \sum_{k=1}^{p} \beta_k + \sum_{i=1}^{n} \xi_i \\ &\text{s.t. } \forall i : \xi_i = \max_{u \neq y_i} l(f_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\beta}}(\boldsymbol{x}_i, y_i) - f_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\beta}}(\boldsymbol{x}_i, u)) \end{aligned} \tag{12}$$

This problem can be solved by iteratively solving $\boldsymbol{\beta}$ with fixed $\boldsymbol{w}$ and $\boldsymbol{b}$ through linear programming, and solving $\boldsymbol{w}$ and $\boldsymbol{b}$ with fixed $\boldsymbol{\beta}$ through a generic SVM solver such as LIBSVM.

To train the SVMs, we employ multi-channel generalized Gaussian kernels with the χ^2 distance, called χ^2 kernel for short. The χ^2 kernel is defined as

$$K(H_i, H_j) = \exp(-\frac{1}{A}\chi^2(H_i, H_j)), \tag{13}$$

where A is the width of the kernel, and χ_2 distance between any two histograms H_i and H_j is defined by:

$$\chi^2(H_i, H_j) = \frac{1}{2}\left(\frac{(H_i(b) - H_j(b))^2}{H_i(b) + H_j(b)}\right), \tag{14}$$

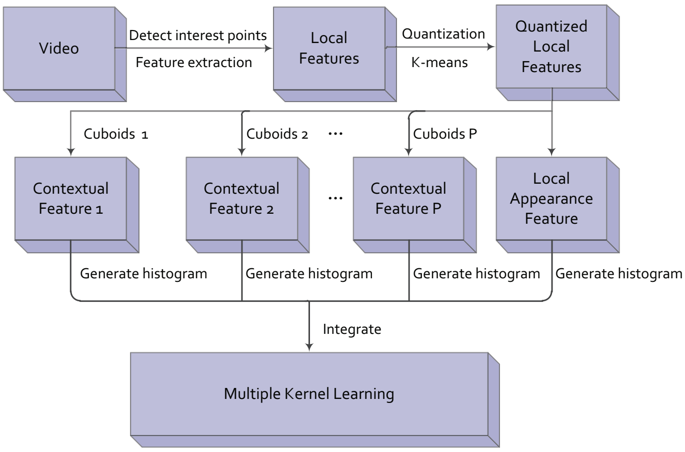

Figure 5. The flowchart of the classification scheme using contextual features

where b indexes over each of the k histogram bins.

4.2. Algorithm Summary

The overview of our algorithm is given in Fig. 5. The classification scheme we consider is a bag of words approach [6]. The interest points are extracted by Harris3D [19]. For each interest point, we use the HOG and HOF features as local appearance features. The K-means clustering algorithm is used to quantize these features into distinct clusters, each of which corresponds to one feature class. These clusters form the codebook for local features.

With these feature classes, we computed P channels the contextual features for each interest point using the methods described in Sec. 3.2. Finally, we cluster the contextual features for each channel i into N_i clusters. These clusters are the codebooks for contextual features.

Based on the codebooks for local features and contextual features, we project local features and each contextual features to the closest codebook element. Then the video is represent by the $P + 1$ histograms of occurrences of codebook elements. Each histogram is called a feature channel.

To perform feature selection in these feature channels, we use a multiclass-MKL with Gaussian-χ^2 kernel for feature selection and classification.

5. Experiments

Our experiments demonstrate the effectiveness of our proposed contextual features for action recognition in a variety of categories.

We evaluate our approach on two benchmark datasets for human activity recognition: the KTH [20] and Activity of Daily Living (ADL) dataset [28].

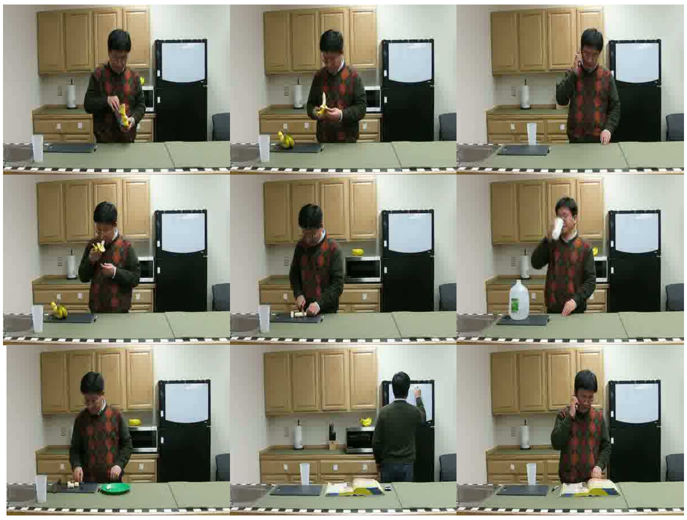

Figure 6. Sample frames from the ADL dataset

5.1. Results on ADL Dataset

The ADL dataset is a high resolution video dataset with 10 different daily-living activities, such as “answering the phone”, “peeling bananas”, and “drinking water”. These activities are each performed three times by five different people for a total number of 150 videos. Some sample frames from ADL dataset are shown in Fig. 6. The activities in this dataset are very similar in appearance, thus they are difficult to be categorized with any single source of information. We will show that the method using our contextual features achieves good performance.

We extract sparse Harris3D points for ADL using the code kindly provided by the authors of [19] with the default parameter setting, and compute Histogram of Gradient (HOG) and Histogram of Flow (HOF) features for each interest point (due to the limitation of the STIP extractor binary code, the videos are subsampled to 400×300). The HOG and HOF features are concatenated to form the local features for each interest point. The size of the codebook for appearance features is fixed as 4000. For contextual feature, we use $W_x, W_y, W_z = (1, 2, 2), (1, 1, 2), (1, 2, 1)$ and $(1, 3, 3)$, with cuboid-size $(40, 40, 20)$. Notice W_x is fixed to be 1 because the horizontal arrangement of interest points does not contain semantic information, as the actions in this dataset can often be interchanged from left-to-right. Each channel of the contextual features is clustered into 1000 classes. For multiclass-MKL, we use the Shogun machine learning toolbox [15]. The penalty parameter is selected by the leave-one-person-out cross-validation.

Method	Accuracy
Messing, *et al.*[28]	89%
Banabbas, *et al.* [2]	81%
Raptis, *et al.*[31]	82.67%
Matikainen, *et al.* [27]	70%
Satkin, *et al.* [33]	80%
STIP interest point	85%
Average among all kernels	94%
Multiple kernel learning	**96%**

Table 1. Performance comparison on ADL dataset with reference methods

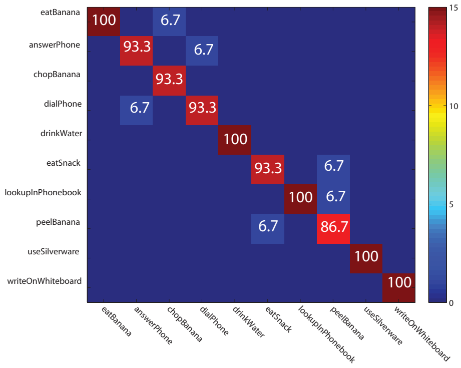

Figure 7. The confusion matrix for ADL dataset

We achieve average accuracy of 96% with multiple kernel learning on the feature channels. This is a very significant improvement to the accuracy compared to the previous work. We also find that learning best kernel combination with MKL is better than simply averaging over all the kernels. Table 1 compare our results to those from previous work. The confusion matrix is shown in Fig. 7. The confusion usually occurs between actions in which the same object exists, such as “answering phone” and “dial the phone”.

To demonstrate the discriminative power of contextual features, we consider two frames extracted from one “eating snack” video and one “eating bananas” video, which are illustrated in the first row of Fig. 8. The yellow circles in Fig. 8 indicate the location of interest points (one picture shows interest points in five consecutive frames). The second and the third row plot the local appearance and contextual feature histogram, respectively. In this case, the local appearance feature histograms are very similar (the sum squared distance between two histograms are 0.2), because interest points in both videos have similar appearance and motion. But their contextual features are quite different, because each of these interest points interacts with different objects. As a result, in some cases, incorporating the contextual feature into classification increases the accuracy of the classification.

5.2. Results on KTH Dataset

The KTH dataset is chosen because of its popularity, although there is not much human-object interaction in this dataset. It contains six action classes, each of which is per-

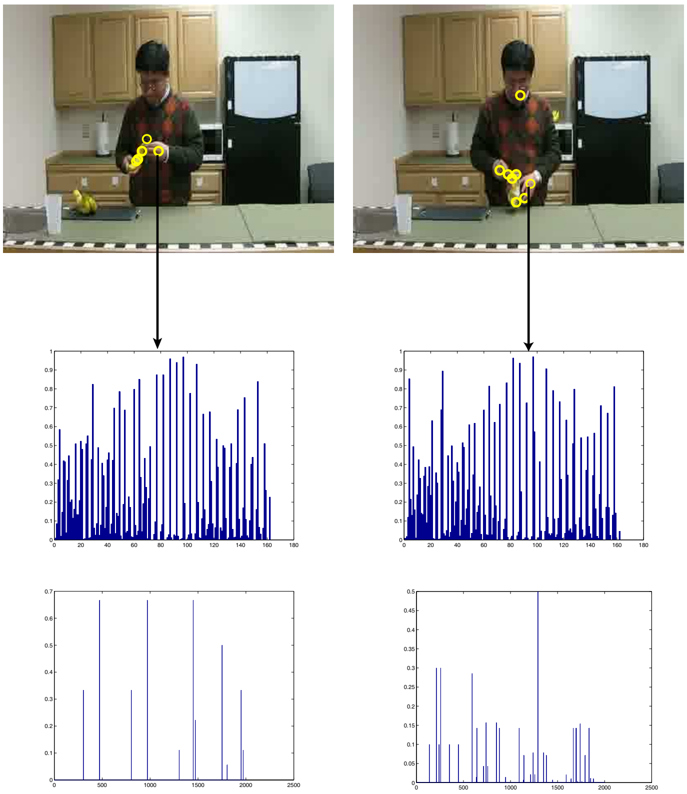

Figure 8. The comparison of local features and contextual features. The top row shows some raw video frames with interest point locations indicated by yellow circles. The second row shows the local feature histogram of the corresponding interest points. The third row contains the contextual feature histogram. In this case, the two local feature histograms are very similar, but their contextual feature histograms are different.

Method	Accuracy
Fathi *et al.* [9]	90.5%
Gilbert *et al.* [12]	89.92%
Lin *et al.* [23]	95.0%
Chen *et al.* [5]	93.4%
Niebles *et al.* [29]	81.5%
Bregonzio *et al.* [3]	93.17%
Laptev *et al.* [20]	91.8%
Kovashka *et al.* [17]	94.53%
Contextual feature	**93.8**%

Table 2. Performance comparison on KTH dataset with reference methods

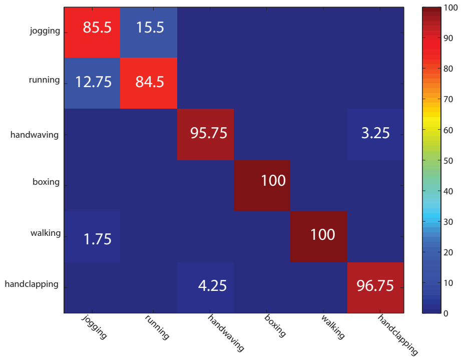

Figure 9. The confusion matrix for KTH dataset

formed in four different scenarios by 25 subjects, resulting in a total of 599 video clips. We follow the split suggested in [20] to split them into 2400 video sequences. Despite uncluttered static background and simplicity of the actions, KTH dataset is relatively difficult because some categories in this dataset are too similar, such as "running" and "jogging".

We choose the same experiments parameter as that in Sec. 5.1. With this setting, we achieve average recognition accuracy of 92.0% without using contextual features, which is comparable to the results in [20], and 93.8% accuracy is obtained with the help of the contextual features. Table 2 compare our results to those from previous work. The performance of our algorithm is comparable to that of the state of art algorithms. Fig. 9 illustrates the confusion matrix for KTH dataset. It can be observed from the confusion matrix that our error mainly concentrates in classifying "jogging" and "running". These actions are very difficult to classify even for human beings, and do not contain many interactions. For other categories, we get better or comparable results.

6. Conclusions and Future Work

We have presented a new type of spatio-temporal contextual feature, based on the density of all features observed in each interest point's spatio-temporal contextual domains. This feature has been validated on two realistic datasets. Our experiments demonstrate the superior performance over the state of art algorithms. In our future work, we intend to develop more efficient learning algorithm to combine the information from different feature channels other than linear combination, and to evaluate our framework using different types of interest point detector and local features.

7. Acknowledgements

This work was supported in part by National Science Foundation grant IIS-0347877, IIS-0916607, and US Army Research Laboratory and the US Army Research Office under grant ARO W911NF-08-1-0504.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(24):509–522, 2002. 3

[2] Y. Benabbas, A. Labl, N. Ihaddadene, and C. Djeraba. Action Recognition Using Direction Models of Motion. In *ICPR*, pages 4295–4298. IEEE, Aug. 2010. 6

[3] M. Bregonzio, S. Gong, and T. Xiang. Recognising Action as Clouds of Space-Time Interest Points. In *CVPR*, 2009. 7

[4] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002. 5

[5] M.-Y. Chen and A. Hauptmann. MoSIFT : Recognizing Human Actions in Surveillance Videos. *Technical Report CMU-CS*, 2009. 1, 2, 7

[6] G. Csurka, C. Bray, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision ECCV*, page 22, 2004. 5

[7] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, pages 886–893. Ieee, 2005. 3

[8] N. Dalal, B. Triggs, and C. Schmid. Human Detection Using Oriented Histograms of Flow and Appearance. In *ECCV*, pages 428–441, 2006. 3

[9] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*,, 2008. 1, 2, 3, 7

[10] R. Filipovych and E. Ribeiro. Recognizing primitive interactions by exploring actor-object states. In *CVPR*, number June, pages 1–7. IEEE, 2008. 3

[11] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, June 2010. 2

[12] A. Gilbert, J. Illingworth, and R. Bowden. Compound Features Mined from Dense Spatio-temporal Corners. In *ECCV*, pages 222–233, 2008. 1, 2, 7

[13] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *ICCV*, pages 925–931. IEEE, 2009. 2, 3

[14] A. Gupta and L. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, number 2, pages 1–8. IEEE, 2007. 3

[15] S. Henschel, G. Ratsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl, and V. Franc. The SHOGUN Machine Learning Toolbox. *JMLR*, 11:1799–1802, 2010. 6

[16] A. Kläser, M. Marszaek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. *BMVC*, pages 995–1004, 2008. 1, 2

[17] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, pages 2046–2053, 2010. 3, 7

[18] D. Kuettel and M. Breitenstein. What's going on? Discovering spatio-temporal dependencies in dynamic scenes. *CVPR*, 2010. 2

[19] I. Laptev. On Space-Time Interest Points. *IJCV*, 64(2-3):107–123, Sept. 2005. 1, 2, 5, 6

[20] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, volume 1, pages 1–8, 2008. 1, 2, 3, 5, 7

[21] Y. J. Lee and K. Grauman. Object-Graphs for Context-Aware Category Discovery. In *CVPR*, 2010. 2

[22] L.-J. Li and L. Fei-Fei. What , where and who ? Classifying events by scene and object recognition. In *ICCV*, 2007. 2

[23] Z. Lin, Z. Jiang, and L. Davis. Recognizing actions by shape-motion prototype trees. In *ICCV*, 2009. 2, 7

[24] F. Lv and R. Nevatia. Single view human action recognition using key pose matching and viterbi path searching. In *CVPR*, pages 1–8, 2007. 2

[25] M. Marszalek and I. Laptev. Actions in context. In *CVPR*, number i, pages 2929–2936, 2009. 2

[26] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, number i. IEEE, June 2009. 2

[27] P. Matikainen, M. Hebert, and R. Sukthankar. Representing Pairwise Spatial and Temporal Relations for Action Recognition. In *ECCV*, pages 508–521, 2010. 3, 6

[28] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, pages 104–111. IEEE, Sept. 2009. 2, 5, 6

[29] J. C. Niebles, H. Wang, and L. Fei-fei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *IJCV*, 2008. 7

[30] H. Ning, W. Xu, Y. Gong, and T. Huang. Latent Pose Estimator for Continuous Action. *ECCV*, pages 419–433, 2008. 2

[31] M. Raptis and S. Soatto. Tracklet Descriptors for Action Modeling and Video Analysis Spatio-temporal Tracklet Descriptors. In *ECCV*, pages 577–590, 2010. 2, 6

[32] M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH: A Spatial-temporal Maximum Average Correlation Height Filter for Action Recognition. *CVPR*, pages 1–8, June 2008. 2

[33] S. Satkin and M. Hebert. Modeling the Temporal Extent of Actions. In *ECCV*, pages 536–548, 2010. 6

[34] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *CVPR*, pages 2004–2011, 2009. 2, 3

[35] Y. Wu and J. Fan. Contextual flow. In *CVPR*, volume 36, pages 33–40. Ieee, June 2009. 3

[36] B. Yao and L. Fei-Fei. Grouplet: a structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 2, 3

[37] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, volume 1, pages 17–24. IEEE, 2010. 3

[38] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *CVPR*, pages 2442–2449. Ieee, June 2009. 2

[39] J. Yuan and Y. Wu. Context-Aware Clustering. In *CVPR*, 2008. 3

[40] A. Zien and C. Ong. Multiclass multiple kernel learning. In *ICML*, volume 1, pages 1191–1198, 2007. 5