Human Action Recognition and Localization in Video using Structured Learning of Local Space-Time Features

Tuan Hue Thi, Jian Zhang
National ICT of Australia & University of New South Wales, NSW, Australia
TuanHue.Thi,Jian.Zhang@nicta.com.au

Li Cheng
TTI at Chicago, Illinois, USA
LiCheng@tti-c.org

Li Wang
Southeast University, Nanjing, China
wang.li.seu.nj@gmail.com

Shinichi Satoh
NII, Tokyo, Japan
satoh@nii.ac.jp

Abstract

This paper presents a unified framework for human action classification and localization in video using structured learning of local space-time features. Each human action class is represented by a set of its own compact set of local patches. In our approach, we first use a discriminative hierarchical Bayesian classifier to select those space-time interest points that are constructive for each particular action. Those concise local features are then passed to a Support Vector Machine with Principal Component Analysis projection for the classification task. Meanwhile, the action localization is done using Dynamic Conditional Random Fields developed to incorporate the spatial and temporal structure constraints of superpixels *extracted around those features. Each* superpixel *in the video is defined by the shape and motion information of its corresponding feature region. Compelling results obtained from experiments on KTH [22], Weizmann [1], HOHA [13] and TRECVid [23] datasets have proven the efficiency and robustness of our framework for the task of human action recognition and localization in video.*

1. Introduction

There are two dominant approaches reported in the literature of action recognition, one is based on the global context of the human shape to infer the actions [1][9][8], while the other looks at how the local key points contribute to the overall action [3][13][4][16][18]. The former group focuses on the whole biological structure of human body, and the human action recognition is based on finding a 3D kinematic model associated with each type of action. The advantage of this approach lies in the fact that, once the model is built, any kind of actions can be inferred from that root, and the analysis result is meaningful and comprehensible by human. However, since input camera source is often in 2D, building a concrete 3D model from these source is not a trivial task, and in fact, video data from realistic context (as seen in Figure 1 for TRECVid dataset) containing large amount of occlusion makes it nearly an impractical task to model the entire body structure. That drawback recently turns computer vision attention to using local features detected at different locations in the video, those features are developed to be invariant to scale, less sensitive to illumination, and can be easily represented as well as analyzed.

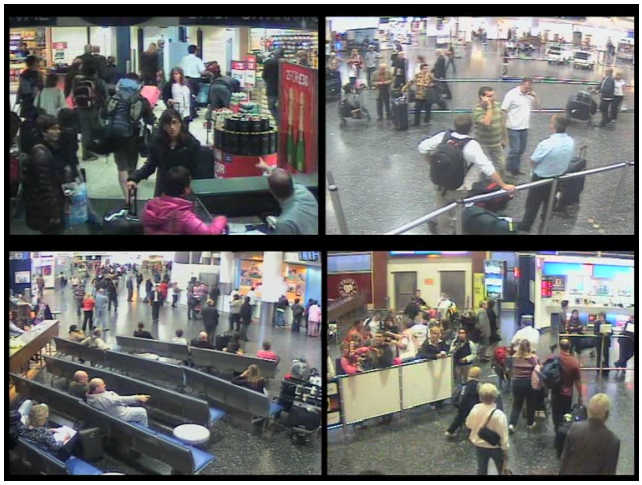


Figure 1. Snapshots from Event Detection task - TRECVid dataset

Adapting similar concepts from the image domain like Harris Corner detector, Scale Invariant Feature Transform (SIFT), many good local feature detection for video have been developed [12][3] to find the the most interesting local regions in the video that can be potential search space

(a) Step 1: STIPs detected as green circles with centers at red points (b) Step 2: HBFS eliminates those irrelevant features in yellow circles (c) Step 3: PCA-SVM decides if this sequence is of class **Embrace** (d) Step 4: CRF weighs features and localizes action by thresholding

Figure 2. Our system framework for Action Classification and Localization, demonstrated on action **Embrace** of TRECVid dataset

for the action. The feature detection step is normally followed by a classification task using different machine learning techniques. The classification stage incorporates the detected local features into different kinds of model, in an attempt to generate the compact representation of high dimensional training data. Common approaches feed the detected local features directly into some kind of discriminative classifier, like Support Vector Machine in [13], bag-of-words based in [3], or shape transformation in [19][17]. Although reported results from those methods demonstrate the validation of those approaches, it is easily noted that the datasets used in those experiments are too simple compared to real world recognition scenarios, which dataset recently made public in [15][14].

Sharing the same view, we tackle human activity recognition in video from the real-world perspective, that is, given a very noisy dataset, our aim is to build a model that can depict the core characteristic of each action class and use them to build a probabilistic model for classifying and localizing their instances in unknown video. In this paper, we will show that it is more effective for the task of action classification, especially in challenging scenarios, if we add an additional step for filtering out noisy local features. We combine classification and localization into a unified framework with four main steps shown in Figure 1, and our main contribution in this work is twofold. Firstly, we extend a Hierarchical Bayesian Feature Selection (HBFS) developed by [2][4] for 2D object detection to the case of 3D Space-Time Interest Point (STIP) [12] action features, as described in Section 2 and 3. Those filtered local features are passed to a non-linear Support Vector Machine with Principle Component Analysis projection (PCA-SVM) [21], described in Section 4, to classify different classes of human action, similar to other traditional approaches. Secondly, we develop a Dynamic Conditional Random Field (DCRF) [11] to weigh each local feature using the combination of its motion-shape descriptor and its neighborhood superpixels, this will be elaborated in Section 5.

2. STIP for Human Action Representation

Visual information of a video $\mathcal{V}$ is defined by a collection of its pixels I, that is $\mathcal{V} \supset I(r, c, t, i)$ with coordinates (r, c, t) *(row, column, time)* and intensity i. We approach video action in an analogous way, decomposing an action $\mathcal{A}$ into local salient patches x, extracted around interest points. Among many local feature choices, we use Space Time Interest Points (STIP) developed from Laptev [12] to represent human action in video, since it yields more informative analysis over the motion in complex background, which is our main goal. The main idea of STIP is to extend Harris interest point detector from 2D image to 3D video, trying to find the point which has significant changes in both directions of space and time [12]. The interest points are detected by searching for the area with high gradient change in shape and motion.

Within each detected local patch x, there can be a range of information that can be embedded from the video domain, and together, all these elements will build up the distinctive characteristics of the underlying action. An local feature x in our design is defined by a feature vector $x(r, c, t, s, z)$, where r, c, and t indicate the geometric position of x, s specifies its scale in region radius, z is the feature description, $z = (hog, hof)$, storing shape and motion information of x as Histogram of oriented Gradients (HoG) and Histogram of oriented Flows (HoF) respectively. Using this representation, the first step in our framework is to extract all STIPs from the video and store in $x \in X$.

Figure 3 shows few result snapshots of STIP extracted from TRECVid dataset at different scale levels.

3. HBFS for concise Action Features

In the current works on human activity analysis, there has been a little number of public dataset that gives the correct annotation of the action class, KTH [22] and Weizmann [1] are probably the two only datasets that have close to complete annotation of when the actions occur in the video shots, Hollywood Human Action (HOHA) [15] is a newly developed dataset trying to include more realistic scenarios, but the annotation is still limited. In fact, video labeling is

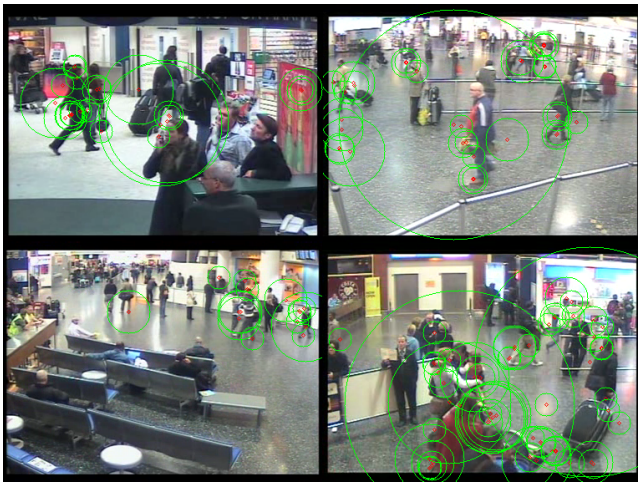


Figure 3. Sample STIPs detected in green circles on TRECVid

much more tedious and time-consuming than the traditional object masking in image recognition. The vast amount of growing video has also brought in the need for a technique that can learn the most representative local features of each action class and be able to catch similar motion pattern in completely unknown environment.

Among many popularly known classification techniques, Bayesian learning approach seems to fit most to our interest of semi-supervised learning task, since it is more flexible in representing the divergence of learning and testing data source, and explicitly shows the link between each hypothesis with its computed score. The core idea of Bayesian approach is to analyze the approximation of the posterior distribution based on multiple trained hypotheses. We extend the Hierarchical Bayesian idea of object recognition in image from Carbonetto *et al*. [2] into human action recognition in video with more constraints on the structure among interest points in both space and time. Each action class will have one classifier trained from its small supervised set, the negative samples are randomly sampled from the pool of all other classes.

For each interest point x_i described as $x(r, c, t, s, z)$ in Section 2, there will be associated a class label $y_i^k \in \{-1, 1\}$. The idea is to build a hierarchical Baysesian classifier model with parameters learned from the limited amount of available training data. Following Carbonetto *et al*. [2], we adopt a sparse kernel machine for classification purpose, with the function between the posterior probability p and probit link Φ defined in Tham *et al*. [25]:

$$p(y_i = 1|x_i, \beta, \gamma) = \Phi(f(x_i, \beta, \gamma)) \tag{1}$$

with f is the regression function

$$f(x_i, \beta, \gamma) = \sum_{k=1}^{N} \beta_k \gamma_k \psi(x_i, x_k) \tag{2}$$

and $\psi(x_i, x_k) = exp(-(x_i - x_k)/\sigma)$, the Gaussian kernel function of x_i with N feature points in the sampling. The two parameters of this classification model are the regression coefficients $\beta \triangleq [\beta_1 \beta_2 ... \beta_N]$ and the feature selection vector $\gamma \triangleq [\gamma_1 \gamma_2 ... \gamma_N]$, $\gamma_k \in \{0, 1\}$, implying the sparsity of this classification [2].

In order to increase the flexibility of the model, we adopt the idea described in Carbonetto *et al*. [10] to assign both parameters β and γ with relevant distributions, respectively β with an inverse Gamma distribution, and γ with Beta distribution. The binary classification of label y_i as shown in [2] is now the calibration of regression function $f(x_i, \beta, \gamma)$ (Equation 2) over zero.

$$y_i = \begin{cases} 1 & \text{if } f(x_i, \beta, \gamma) > 0 \\ -1 & \text{otherwise} \end{cases} \tag{3}$$

The discriminative classification now becomes calculation of the probability of a new point x' based on training data $\{x, y_k\}$, and model parameters $\theta = \{\beta, \gamma\}$

$$p(y'|x', x, y_k) = \int p(y'|x', \theta) p(\theta|x, y^k) d\theta \tag{4}$$

The computation of equation 4 is clearly explained in [2] using Markov Chain Monte Carlo sampling in addition with a blocked Gibbs sampler as advised by Tham *et al*. in [25]. Figure 4 shows few snapshot results of action *PersonRuns* in TRECVid, there are still different false labeling because of the noisy background, but essentially the event region is covered.

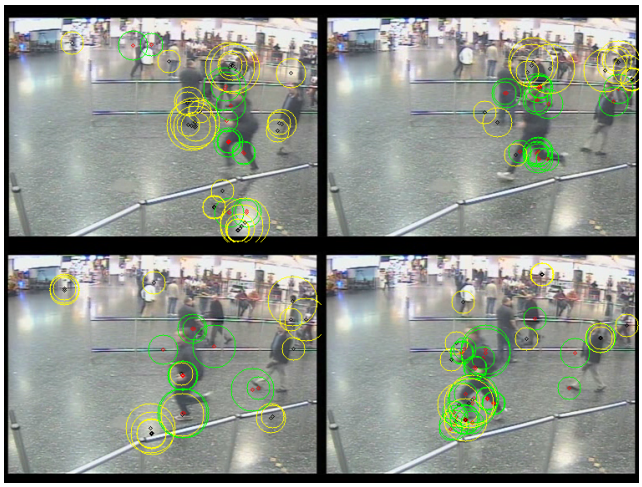


Figure 4. HBFS labels green circles as relevant features for action **PersonRuns**, yellow circles for noise which will be removed

At this stage, we have represented an action instance in video using a only the finest set of local features x which has discriminative feature label $y = 1$. This additional feature

selection stage will be quantitatively evaluated in Section 6.2.

In most literatures of this bag-of-feature approach [16][2], people often use the feature labeling results independently and directly to classify the context. In our approach, we see that the contributing correlation among different local features is important to represent a human action as a whole. Therefore, we apply an extra discriminative classifier for this purpose, using a non-linear PCA-SVM, similar approaches to Laptev *et al*. in [13].

4. PCA-SVM for Action Classification

After the feature labeling task, each video shot can be seen as a sparse set of all event points $i(r, c, t, s, d, l)$ with label $l = 1$ indicating all these points belong to this action class of interest. Using the Radial Basis Function as the Support Vector Machine kernel [21] $k(x_i, x_j) = exp(-\gamma \|x_i - x_j^2\|)$, for $\gamma > 0$, we generate a classifier model from the supervised part, and use it to classify unknown event shots. The task of action classification is done using one-against-all, that is, when one action is used to build the classifier, all instances of other classes are considered as negative samples. Figure 5 shows the binary classification results of ObjectPut action classifier.

Figure 5. Using action model **ObjectPut**, PCA-SVM classifies the left video shot (blue text with (+) sign) as a positive instance, and the right shot (red text with (-) sign) as a negative

5. DCRF for Action Localization

Often in the image object recognition task, objects are detected and localized at certain bounding boxes which are helpful to show the exact object location, and also, can be used as a ground-truth data for further detection. However, in video processing domain, the concept of human activity or human event is rather abstract and loosely defined, especially for those videos obtained from the web [14] or real world surveillance scenarios TRECVid [23], the automatic retrieval of event regions is very essential and helpful for the activity analysis society.

In the classification task described in the previous section, local features are independently projected and used to find the support vectors, those best discriminate one action class from others. Meanwhile, with the challenging task of action localization, the aim is to retrieve only the features that directly construct the action regions. In order to decide which features should be used to construct the action rectangular cuboid, we introduce a concept of feature relevancy weight $w \in [0, 1]$ represents the relevance of each feature with the action. In our approach, we call the action cuboid Integral Volume, which basically is a bounding cuboid of all features $x|w(x) > \eta$, with η is the relevancy weight threshold of features, distinctive for each action class. Estimation of w is done by formalizing the two observations about features of a common human action. The first observation is spatial dependency, neighboring features $x, y|y \in N_x, d(x, y) < \tau)$ are likely to have similar contribution weight to an action region, here N_x is the spatial neighborhood set of x, d is the normalized Euclidean distance and τ is the neighborhood distance threshold. The second observation is temporal dependency, the action regions in adjacent frames normally do not have large variance in size and location, in other words, same features across time dimension x_k and x_{k+1} tend to have similar weights, here k indicates time frame.

By encoding spatial and temporal dependencies of neighboring features into the selection process, we have converted the localization task into structured learning with latent variables. The hidden parameter in our model is the feature weight w, and the structured dependencies are decomposed into spatial and temporal constraints. Among many structured learning techniques, Conditional Random Fields (CRF) [11] are most appealing to our case of dependent sparse local features. For the task of object localization in images, Carbonetto *et al*. [2] had successfully applied a standard CRF to model spatial constraints. Specifically for our action localization task with additional temporal constraints, we employ the approach in [26] to develop a Dynamic Conditional Random Fields (DCRF) with an extra temporal constraint. Wang and Ji in [26] uses DCRF for the problem of object segmentation from video with dense features, which are in fact all the pixels in the video. In our case, we use sparse local *superpixels* x, the 3D cuboid extracted around STIP, as the feature observations, shown as small green rectangles in Figure 6(a), to find the bounding cuboid of the action instance in the video shot.

Formally, we denote z as the feature observation, $z = (hog, hof)$ in our case for Histogram of oriented Gradients hog and Histogram of oriented Flows hof representing feature shape and motion respectively. The feature weight w is now a random field globally conditioned on z. Using the Hammersley-Clifford theorem and considering only one-pixel and two-pixel potentials, we now can represent the posterior probability $p(w_k|z_{1:k})$ of the feature weight given

z by a Gibbs distribution as

$$p(w_k|z_{1:k}) \propto exp\{-\sum_{x\in X}[\varphi_x(w_k(x)|z_{1:k})+ \sum_{y\in N_x}\varphi_{x,y}(w_k(x),w_k(y)|z_{1:k})]\} \quad (5)$$

In this equation, X is the local feature domain, $z_{1:k}$ is the observed feature sequence up to time k, $\varphi_x(w_k(x)|z_{1:k})$ is the one-pixel potential function for each *superpixel* x, $\varphi_{x,y}(w_k(x),w_k(y)|z_{1:k})$ is the two-pixel potential function representing the spatial constraint between a pair of two neighboring features. The temporal constraint is formulated as two potentials $\varphi_x(w_{k+1}(x)|w_k(N'_x))$ and $\varphi_{x,y}(w_{k+1}(x),w_{k+1}(y))$ and encoded in the state transition probability as developed by Wang and Ji

$$p(w_{k+1}|w_k) \propto exp\{-\sum_{x\in X}[\varphi_x(w_{k+1}(x)|w_k(N'_x))+ \sum_{y\in N_x}\varphi_{x,y}(w_{k+1}(x),w_{k+1}(y))]\} \quad (6)$$

with N'_x is the temporal neighborhood set of x, containing neighbors of x in the adjacent state. Apart from the posterior and state transition function, the likelihood function $p(w_k|z_k)$ is also derived similarly to [26] as

$$p(z_k|w_k) \propto exp\{-\sum_{x\in X}[\varphi_x(z_k|w_k(x))+ \sum_{y\in N_x}\varphi_{x,y}(z_k(x),z_k(y)|w_k(x),w_k(y))]\} \quad (7)$$

where $\varphi_x(z_k|w_k(x))$ and $\varphi_{x,y}(z_k(x),z_k(y)|w_k(x),w_k(y))$ are similarly the one and two-pixel potentials representing the spatial constraints of shape-motion observation and feature weights. Since motion and shape are retrieved independently, the likelihood function can be further decomposed to

$$\begin{aligned} p(z_k|w_k) &= p(hog_k, hof_k)|w_k) \\ &= p(hog_k|w_k)p(hof_k|w_k) \end{aligned} \quad (8)$$

The optimization process is carried out similarly to the segmentation sampling described in [26], by approximating the mean field probability $q_x(w_k(x)|z_{1:k})$

$$p(w_k|z_{1:k}) \approx \prod_{x\in X} q_x(w_k(x)|z_{1:k}) \quad (9)$$

$$\widehat{w}_k(x) = \arg\max_e q_x(w_k(x)=e|z_{1:k}) \quad (10)$$

where e is the initialization value, $q_x(w_o(x)=e)$, and is set to 0.5 for all *superpixel* x in our case. The calculated $\widehat{w}_k(x)$ is the final feature weight of all *superpixels* in the video shot, which will be passed through the weight filter η as described previously. The final *Integral Volume* is caluclated as the as the approximate bounding rectangular cuboid that contains all those high weight features. Figure 6 illustrates the localization results using DCRF for an instance of action Embrace from TRECVid dataset.

(a) Extracted rectangular cuboids at STIP location. Note that all these cuboids already selected by the previous HBFS

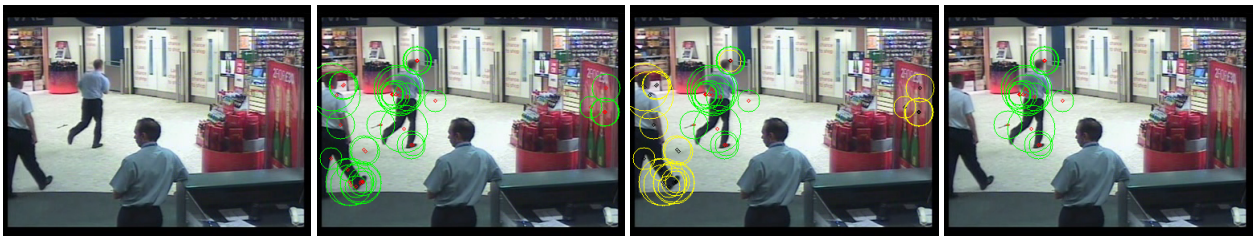
(b) DCRF results, grayscale color of the cuboids represent different feature weights, Integral Volume is drawn in green rectangle

Figure 6. Feature Relevancy Weighting using DCRF

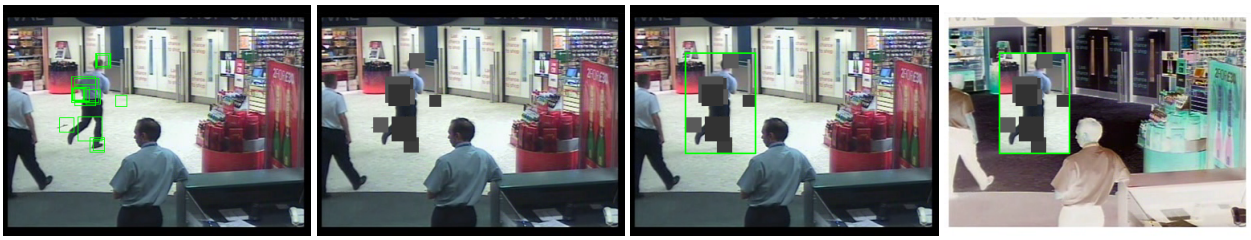
6. Experimental Results

6.1. Dataset selection and Experiment Setup

In order to evaluate the performance of our proposed approach, we run action classification and localization tasks on four main datasets which represent different characteristics of testing scenarios. For the fundamental and complete set of human action recognition, we choose KTH [22] and Weizmann [1] dataset. KTH has about 2400 grayscale video shots with 6 actions: *boxing*, *handwaving*, *handclapping*, *jogging*, *running*, *walking*, performed by 25 persons under 4 different contexts and subdivided into 4 intervals. Weizmann has about 90 colored video shots with 10 actions: *bend*, *jack*, *jump*, *pjump*, *run*, *side*, *skip*, *wave1*, *wave2*, *walk*, performed by 9 persons. For the realistic set, we choose Hollywood Human Action HOHA1 dataset [13], which we can find more publicly reported works than its descender HOHA2 [15], HOHA1 contains 8 action classes, namely *AnswerPhone*, *GetOutCar*, *HandShake*, *HugPerson*, *Kiss*, *SitDown*, *SitUp*, and *StandUp*, distributed in around 250 training and testing samples. For a even more challenging scenario of action recognition in surveillance video, we pick TRECVid Event Detection Development set [23], recorded from 5 cameras at Gatwick airport in the United Kingdom. Using the provided annotation file together with 20 video shots recorded in 4 different days from 4 main cameras, excluding camera 4 looking only at the elevator and has very little action, we extract all the associated samples to build a dataset of 5584 action samples of 8 different action events, namely *CellToEar* 398 shots, *Embrace* 449 shots, *ObjectPut* 984 shots, *OpposingFlow* 15 shots, *PeopleMeet* 1246 shots, *PeopleSplitUp* 761 shots, *Person-*



(a) Stage 1: Original Frame (b) Stage 2: STIP (c) Stage 3: HBFS (d) Stage 4: SVM Classification



(e) Stage 5: Superpixels (f) Stage 6: CRF Feature Weighting (g) Stage 7: Integral volume (h) Stage 8: Action Localization

Figure 7. Detailed steps for recognizing action *PersonRuns* from TRECVid Event Detection Track

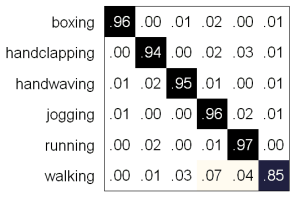
Runs 281 shots, and *Pointing* 1452 shots.

Figure 7 shows 8 detailed output stages of our action classification and localization framework. The classification and localization results obtained from running our system on the 4 datasets are then used to evaluate the performance of our system compared with state-of-the-arts in the field.

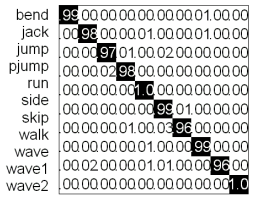
6.2. Action Classification

In order to provide a fair comparison with other approaches, the task of action classification on each dataset is performed with different amount of training and testing. On KTH, we use *2/3 Split*, that is, 1800 shots for training and 900 shots for testing, dividing based on person and context variation. On Weizmann, we use *Leave-One-Out* scheme to train and test all sequences. On HOHA, we use the same number of training and testing that Laptev *et al*. used in [13] and for TrecVID, we use *2/3 Split* for each action class.

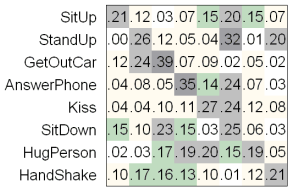
The results obtained from running the classifier on three datasets are shown in 4 confusion matrices of Figure 8. We can easily see that our system performs much better on KTH (93.83%) and Weizmann (98.2%), which is reasonable since those two dataset does not have occlusion and only one actor is visible at a time. We use the average classification accuracy to compare with reported state-of-the-art systems in Table 1. We can see that on KTH and HOHA, our system outperforms [13] (91.80% on KTH, and 18.88% on HOHA) and [3] (81.20% on KTH, and 6% on HOHA) which both use STIP as local features. This has proved the effectiveness of the additional HBFS step that we added before the discriminative classification. Finally, it is noticed that the performance on TRECVid dataset is a quite low,



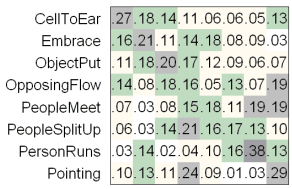
(a) KTH average rate 93.83%



(b) Weizmann average rate 98.2%



(c) HOHA average rate 26.63%



(d) TRECVid average rate 23.25%

Figure 8. Confusion Matrix for Action Classification Performance

nearly 23.25% in average, showing the challenging characteristic of this dataset with frequent occlusion and low resolution video quality.

6.3. Action Localization

Action localization from video shot is carried out as described in Section 5. The evaluation is carried out on those true positive video samples from the 4 datasets.

The labeling task for groundtruth localization data is highly time-consuming, so we only select a portion of each dataset to quantitatively test the proposed localization ap-

(a) Step 1: STIP Detection (b) Step 2: HBFS (c) Step 3: PCA-SVM (d) Step 4: CRF Weighting (e) Step 5: Localization

Figure 9. Classification and Localization of Action **CellToEar** in *TRECVid Camera 1*

(a) Step 1: STIP Detection (b) Step 2: HBFS (c) Step 3: PCA-SVM (d) Step 4: CRF Weighting (e) Step 5: Localization

Figure 10. Classification and Localization of Action **CellToEar** in *TRECVid Camera 2*

(a) Step 1: STIP Detection (b) Step 2: HBFS (c) Step 3: PCA-SVM (d) Step 4: CRF Weighting (e) Step 5: Localization

Figure 11. Classification and Localization of Action **CellToEar** in *TRECVid Camera 3*

(a) Step 1: STIP Detection (b) Step 2: HBFS (c) Step 3: PCA-SVM (d) Step 4: CRF Weighting (e) Step 5: Localization

Figure 12. Classification and Localization of Action **CellToEar** in *TRECVid Camera 5*

proach. With KTH and Weizmann, the scenario only contains one actor in a video scene, so we run an automatic motion detection with filters to find the human area, and consider the bounding mask as the groundtruth data. Meanwhile, for HOHA and TRECVid, we develop a video labeling tool to segment region of events in both frame number and action bounding box. The localization is evaluated as finding the overlapping proportion of calculated bounding box (on positively classified video shots) with the groundtruth, an overlapping region higher than 50% is said to be a correct localization.

Table 2 summarizes the localization results together with number of evaluated video shots. Similar to classification task, localization on KTH and Weizmann provide higher performance, especially on Weizmann with 100%. Nevertheless, localization on HOHA and TREVid are actually promising providing that those video shots are normally full of noise. These results indicate the robust performance of our proposed localization framework using DCRF, and can be further developed to use as a semi-supervised labeling tool for action recognition. Figure 9 to 12 show good localization output snapshots for CellToEar action in TRECVid dataset, recorded on 4 different cameras.

7. Conclusion

In this paper, we have presented a unified approach towards solving the human action recognition and localiza-

Approach	KTH	Weizmann
Ours	***93.83***	***98.2***
Sun and Hauptmann [24]	94	97.8
Grundmann et al. [6]	93.52	96.39
Laptev et al. [13]	91.80	*
Jhuang et al. [7]	91.70	98.8
Wang and Mori [27]	91.17	98.33
Fathi and Mori [5]	90.50	100
Rapantzikos et al. [19]	88.30	*
Jiang et al. [8]	84.40	*
Niebles et al. [16]	81.50	72.8
Dollar et al. [3]	81.20	*
Ke et al. [9]	80.90	*
Schuldt et al. [22]	71.70	*

Table 1. Human Action Classification Performance of state-of-the-art approaches on KTH and Weizmann

Dataset	Evaluated Shots	Avg. Accuracy
KTH	400	96.25
Weizmann	80	100
HOHA	120	77.5
TRECVid	600	71.83

Table 2. Action Localization on selected subsets of 4 datasets

tion together in one framework. By introducing an additional discriminative feature selection HBFS step, we have greatly improve the overall recognition over traditional approach. In addition, with an application of DCRF, the challenging task of action localization can now be tackled and evaluated statistically. Results on both research and real-world datasets have shown that our proposed approach is highly competitive with the state-of-the-art approaches, as well as promising for practical applications in the case of human activity analysis in video.

Acknowledgement

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, 2005. 1, 2, 5

[2] P. Carbonetto, G. Dorko, C. Schmid, H. Kuck, and N. de Freitas. Learning to recognize objects with little supervision. *IJCV*, 77(1-3):219–237, May 2008. 2, 3, 4

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *PETS*, 2005. 1, 2, 6, 8

[4] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002. 1, 2

[5] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*, 2008. 8

[6] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007. 8

[7] H. Jiang, M. Drew, and Z. Li. Successive convex matching for action detection. In *CVPR*, 2006. 1, 8

[8] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, 2007. 1, 8

[9] H. Kuck, P. Carbonetto, and N. de Freitas. A constrained semi-supervised learning approach to data association. In *ECCV*, 2004. 3

[10] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001. 2, 4

[11] I. Laptev. On space-time interest points. *IJCV*, (2-3):107–123, September 2005. 1, 2

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1, 2, 4, 5, 6, 8

[13] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos 'in the wild'. In *CVPR*, 2009. 2, 4

[14] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2, 5

[15] J. Niebles, H. Wang, and F. Li. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3), September 2008. 1, 4, 8

[16] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal salient points for visual recognition of human actions. *SMC*, 36(3):710–719, June 2006. 2

[17] D. Ramanan, D. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *PAMI*, 29(1):65–81, 2007. 1

[18] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In *CVPR*, 2009. 2, 8

[19] K. Schindler and L. Van Gool. Action snippets: How many frames does human action recognition require. In *CVPR*, 2008. 8

[20] B. Scholkopf, A. Smola, and K. Muller. Kernel principal component analysis. *ICANN*, pages 583–588, 1997. 2, 4

[21] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004. 1, 2, 5, 8

[22] A. F. Smeaton, P. Over, and W. Kraai. Evaluation campaigns and trecvid. In *MIR*, 2006. 1, 4, 5

[23] X. Sun and M. Hauptmann. Action recognition via local descriptors and holistic features. In *CVPR*, 2009. 8

[24] S.-S. Tham, A. Doucet, and K. Ramamohanarao. Sparse bayesian learning for regression and classification using markov chain monte carlo. In *ICML*, 2002. 3

[25] Y. Wang and Q. Ji. A dynamic conditional random field model for object segmentation in image sequences. In *TPAMI*, 2006. 4, 5

[26] Y. Wang and G. Mori. Human action recognition by semi-latent topic models. *PAMI*, 31(10):1762–1774, 2009. 8