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.

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
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 \( V \) is defined by a collection of its pixels \( I \), that is \( V \supseteq I(r,c,t,i) \) with coordinates \((r,c,t)\) \((\text{row, column, time})\) and intensity \( i \). We approach video action in an analogous way, decomposing an action \( A \) 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, \( z = (\text{hog, hof}) \), 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
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 constrains 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 $y^{x_i} ∈ \{−1, 1\}$. The idea is to build a hierarchical Bayesian 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]:

$$p(y_i = 1|x_i, β, γ) = Φ(f(x_i, β, γ))$$ (1)

where $f$ is the regression function

$$f(x_i, β, γ) = \sum_{k=1}^{N} β_k γ_k ψ(x_i, x_k)$$ (2)

and $ψ(x_i, x_k) = exp(-(x_i - x_k)/σ)$, 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 $β ≜ [β_1, β_2, ..., β_N]$ and the feature selection vector $γ ≜ [γ_1 γ_2 ... γ_N]$, $γ_k ∈ \{0, 1\}$, 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 $f(x_i, β, γ)$ (Equation 2) over zero.

$$y_i = \begin{cases} 1 & \text{if } f(x_i, β, γ) > 0 \\ -1 & \text{otherwise} \end{cases}$$ (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 $θ = \{β, γ\}$

$$p(y'|x', x, y_k) = \int p(y'|x', θ)p(θ|x, y_k)dθ$$ (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.

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 know event shots. The task of action classification is done by formalizing the two observations about features of a common human action. The first observation is spatial dependency, neighboring features \(x, y\) with \(y \in N_x, d(x, y) < \tau\) 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 \(\tau\) 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 can now represent the posterior probability \(p(w_k|z_{1:k})\) of the feature weight given...
the mean field probability as

\[
p(w_k|z_{1:k}) \propto \exp\{- \sum_{x \in X} \phi_x(w_k(x)|z_{1:k}) + \sum_{y \in N_x} \phi_{x,y}(w_k(x), w_k(y)|z_{1:k})\} \tag{5}
\]

with \(N_x^r\) 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

\[
p(z_k|w_k) \propto \exp\{- \sum_{x \in X} \phi_x(z_k|w_k(x)) + \sum_{y \in N_x} \phi_{x,y}(z_k(x), z_k(y)|w_k(x), w_k(y))\} \tag{7}
\]

where \(\phi_x(z_k|w_k(x))\) and \(\phi_{x,y}(z_k(x), z_k(y)|w_k(x), w_k(y))\) 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

\[
p(z_k|w_k) = p(hog_k, hof_k|w_k) \\
= p(hog_k|w_k)p(hof_k|w_k) \tag{8}
\]

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

\[
p(w_k|z_{1:k}) \approx \prod_{x \in X} q_x(w_k(x)|z_{1:k}) \tag{9}
\]

\[
\hat{w}_k(x) = \arg \max_e q_x(w_k(x) = e|z_{1:k}) \tag{10}
\]

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 \(\hat{w}_k(x)\) is the final feature weight of all superpixels in the video shot, which will be passed through the weight filter \(\eta\) as described previously. The final Integral Volume is calculated 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.

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

(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

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, 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-
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 TRECvid 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-
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


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

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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Evaluated Shots</th>
<th>Avg. Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KTH</td>
<td>400</td>
<td>96.25</td>
</tr>
<tr>
<td>Weizmann</td>
<td>80</td>
<td>100</td>
</tr>
<tr>
<td>HOHA</td>
<td>120</td>
<td>77.5</td>
</tr>
<tr>
<td>TRECVID</td>
<td>600</td>
<td>71.83</td>
</tr>
</tbody>
</table>

Table 2. Action Localization on selected subsets of 4 datasets