# Probabilistic and Max-margin structured learning in Human Action Recognition

Tuan Hue Thi[12] Li Cheng[3] Jian Zhang[12]
Li Wang[4] Shinichi Satoh[5]

[1] University of New South Wales, Australia
[2] National ICT Australia
[3] Bioinformatics Institute, A*STAR, Singapore
[4] Nanjing Forest Univeristy, China
[5] National Institute of Informatics, Japan

**Technical Report**
**UNSW-CSE-TR-1110**
**June 2011**


School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

**Abstract**

Human action recognition is a promising yet non-trivial computer vision field with many potential applications. Current advances in bag-of-feature approaches have brought significant insights into recognizing human actions for various practical purposes. It is, however, a common practice in literature to consider a set of local feature descriptors with uniform contributions. This assumption has been shown to be oversimplified, which limit these works from robust deployments in real-life video content retrieval. In this work, we propose and show that, by taking into account global configuration of local features, we can greatly improve the recognition performance. A novel feature selection process is also devised with the help of Sparse Hierarchical Bayes Filter, an additional process to boost the traditional *bag-of-feature* learning. We further introduce the usage of structured learning for the problem of human action recognition. That is, by representing one human action as a complex set of local features, a set of feature functions can be utilized to discriminatively infer the structured output for action classification and action localization. In particular, we tackle the problem of action localization in video using structured learning and we compare two two options: One is Dynamic Conditional Random Field from probabilistic principle; The other is Structured Support Vector Machine from max-margin principle. We evaluate our modular classification-localization framework on various testbeds, where the proposed framework is demonstrated by its competitive performance comparing with the state-of-the-art methods.

# 1 Introduction

Human action recognition has wide range of applications in different areas, including human computer interaction, public surveillance, and multimedia content retrieval. High level semantic information obtained from action recognition can also be directly applicable to various tasks, such as robotics, security, entertainment, and bioinfomatics analysis. There are two main challenges in human action analysis, namely large visual variation, and expensive computational learning-inference. **Visual variation** is introduced by different scene backgrounds, structure of human bodies and human actions. *Background variation* is undoubtedly a common characteristic of videos recorded these days, it occurs mostly due to illumination change, moving camera and partial occlusion. Meanwhile, *human body variation* is caused due to it 3D non-trivial kinematic structure that can be projected in different ways onto 2D images under different recording perspectives, for example, a frontal human image can be totally different from a human image viewed from the top and side. Lastly, semantic classification of human actions is normally too broad to include a vague meaning of different *action instances*; people 'running' at different speed, 'walking' with different styles. These variation challenges require complex modeling techniques to learn and do inference, which in turns, lead to the second obstacle of **intractable computation**. A robust and generic action analysis system would have to include significant number of feature parameters, each of which is involved with large degree of freedoms.

Current approaches in action recognition try to overcome these two previously mentioned challenges by detecting and learning salient visual features of human actions using simplified training and inferring techniques. Simplification is done at feature level where a video is transformed into a single vector of either shape-motion gradients or quantized histogram of visual words. The learning is also commonly limited to binary detection of action existence using traditional supervised approaches from object detection in images. In this work, we will show that not all local features, which detected in the same fashion across action instances, are useful for the task of action classification. In fact, with bag-of-feature learning, feature existence contributes mostly to the prediction outcome, and dominant but irrelevant features might completely change model behavior. An example would be seen in 'AnswerPhone' action, where someone might be walking and talking on the phone, many features might be detected for the walking patterns while the main action of picking up the phone around small arm area only appears in a short period of time. This common drawback can be efficiently solved using an additional feature selection process in which correlation between local features and action class is learned from training data to learn only those most contributive to the classification process. For the typical characteristic of local action features in video, we will introduce *Hierarchical Bayesian Feature Selection* to produce a sparse subset of discriminative features from the input feature set. In addition, we will also formally deal with the challenging problem of action localization, in which the 3D bounding volume of action instance is calculated. This can be seen as a typical structured learning problem, where input domain is sparse set of local features containing hidden interaction, and output domain is a random field of relevancy weights. This localization task is an intractable combinatorial optimization problem where feature space is exponentially large. For this, we will introduce and compare

two efficient structured learning techniques, namely *Dynamic Conditional Random Fields*, and *Structured Support Vector Machine*. These two methods not only make learning large structured sets possible, but also incorporate efficiently hidden constraints of local features on both spatial and temporal domains.

## 1.1 Related work

There are different ways to categorize current approaches in human action recognition. In this paper, we will use a feature-oriented perspective to group closely related works into structure-constrained features and orderless local-features.

**Structure-constrained features** This approach relies on two characteristics of human action, namely rigorous human body structure and temporal tracking. The first common example of this type is the holistic feature where the whole human body is taken into account, and motion field of body movement will be extracted to form the action features. 'Motion history image' (MHI) from Bobick and Davis [2001] is one of the earliest reported holistic feature, in which motion field is concatenated in chronological ascending order, and learning can be effectively done using different moment features. Similar representation of MHI can also be seen in 'motion history volumes' work from Weinland et al. [2006] where multiple cameras are used to synthesize motion field. While desirable invariance and performance can be achieved with this approach, it is apparent that it has limited practical applications. Other holistic approaches can also be found in Efros et al. [2003], Yilmaz and Shah [2005], Zelnik-Manor and Irani [2001], and Ke et al. [2007a] where 'spatio-temporal volumes' of the human body are used to extract global shape contours and motion gradients to search for similar global patterns across action instances. Structure-constrained features also include works that use tracking of a-priori body landmarks to calculate features. Some of the notable works are described in Sigal et al. [2004], Ramanan et al. [2007] and Moon and Chellappa [2008] where decomposable articulated human parts are tracked, or in Abdelkader et al. [2008], Abdelkader et al. [2008], Stenger et al. [2006], and Guo and Qian [2008] with the use of specific body landmarks like torso, legs and arms.

All these structure-constrained features have in common the strength of being interpretable at high semantic level. On the other hand, the main drawback of this group of features relies on two context assumptions, one is the known a-priori model of the human structure, which generates badly and usually unsatisfactory in realistic video analysis. Secondly, these features rely on reliable tracking of pre-defined rigorous body parts, which often fails in cases where large environment variation or occlusion occur, an example of this can be seen in Figure 1.1 from TRECVid dataset.

**Orderless local-features** Meanwhile, this approach detects local features using a set of response filters, which typical aim to detect local salient patterns in human shape and motion. Regardless of the feature source, all that meet saliency criteria will be considered to contribute on the human action. This feature scheme was originally proposed to tackle the drawback of structure-constrained features by considering only a much smaller feature space at sparse scale. It usually yields less informative recognition information of human actions, but is proved to work reasonably well under various conditions including

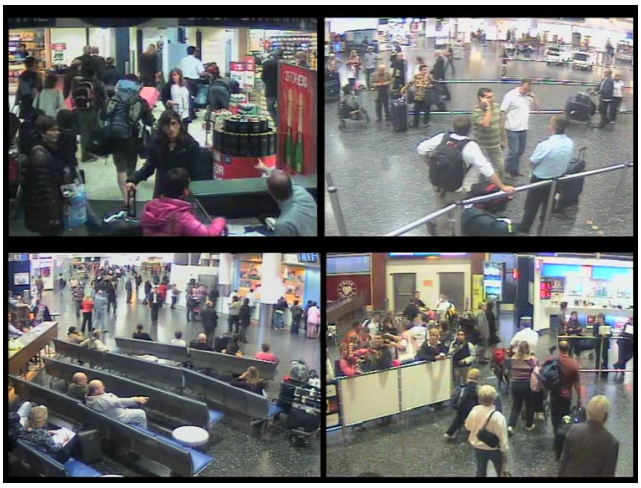


Figure 1.1: Snapshots from *Event Detection* track - TRECVid dataset.

occlusion and cluttered background. Recently, different local saliency response filters have been proposed. Oikonomopoulos et al. [2006] extend saliency point detector from Kadir and Brady [2003] into entropy-based spatiotemporal salient point, Fathi and Mori [2008], Ahmad and Lee [2008], and Shechtman and Irani [2005] use correlation of action templates to look for local salient patches from raw videos. Among all these approaches in this category, the two most commonly used local features are 'space-time interest point' (STIP) from Laptev [2005] and 'space-time cuboid' from Dollar et al. [2005a]. While STIP is developed as an space-time extension of Harris corner detector Harris and Stephens [1988] to detect high variation on both spatial and temporal direction, cuboid is taking into account only local maxima in spatial directions and look for denser sampling of spatio-temporal volume. Various works have been reported that produce compelling results using STIP and cuboids in the bag-of-feature framework. Notable works are in local Support Vector Machine approach of Schuldt et al. [2004a], unsupervised probabilistic topic modeling of cuboids in Niebles et al. [2008a], and weakly supervised learning of local features using 'implicit motion-shape model' (ISM) in Thi et al. [2010a] Thi et al. [2010b]. The main drawback of learning orderless local-feature using *bag-of-feature* approach might be attributed to its negligence of the spatial and temporal structure, which often produces biased learning and in turns, during inference a number of the detected local features are often irrelevant to current action as mentioned previously.

## 1.2 Overview of the framework

In this work, we first introduce an effective solution to local *bag-of-feature* classification using an additional feature selection step based on discriminative training of structured inputs. This feature selection step learns those features that are mostly representative to each action class, and produces filter decision on novel local feature set. Secondly, we use structured learning to solve the task of action localization. There are very few reported works for action localiza-

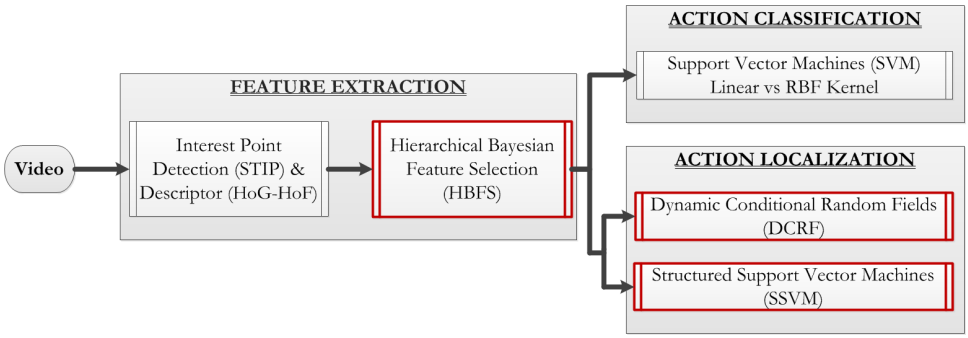
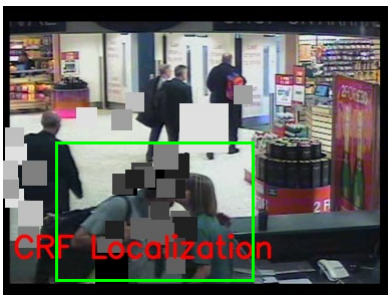



Figure 1.2: The proposed modular framework for action classification and localization using local features. Modules in RED are our main contributions in this work.


(a) Step 1: STIP detected as green circles with centers at red points.


(b) Step 2: HBFS eliminates those irrelevant features in *yellow circles*.


(c) Step 3: SVM decides if this sequence is of class **Embrace**.




(d) Step 4: DCRF weighs features and localizes action by thresholding.

Figure 1.3: Our system snapshot for action classification and localization, demonstrated on action **Embrace** of TRECVid dataset.

tion, this task is usually coupled with classification process where all selected space-time features of a positive prediction will be used to generate the localization boundary, and shown to work on simplified dataset like KTH Schuldt et al. [2004a] and Weizmann Blank et al. [2005], as in Yuan et al. [2009], Niebles et al. [2008b], Oikonomopoulos et al. [2010], and Alexander [2004]. In this paper, we will show that action localization is an essential task for human action

analysis, especially in those situation where background is highly complex with various cluttered scenario and distraction, as in HOHA Laptev et al. [2008a] and TRECVid Smeaton et al. [2006].

We incorporate the dual tasks of action classification and localization into a unified framework, as illustrated in Figure 1.2. This system is most related to one described in our previously published work of Thi et al. [2010c]. We analyze human actions using three main modules, **feature extraction**, which contains Interest Point Detection - *Space Time Interest Point* (STIP) and Descriptor - *Histogram of oriented Gradients and Flows* (HoG-HoF) as well as *Hierarchical Bayesian Feature Selection* (HBFS), will be detailed in Section 2, **action classification** using standard visual word quantization of *bag-of-feature* approach with *Linear* and $\chi^2$ kernels of *Binary Support Vector Machine* (SVM) Scholkopf et al. [1997], detailed in Section 3, and lastly the task of **action localization** will be tackled using two structured learning approaches, namely *Dynamic Conditional Random Fields* (DCRF) and *Structured Support Vector Machines* (SSVM), described in Section 4. Section 5 will show the effectiveness of those introduced structured learning by empirical results of the proposed framework against traditional approaches on four human action datasets, Figure 5.1 shows one snapshot of our proposed system using HBFS on STIP for feature extraction, $\chi^2$ SVM for action classification and DCRF for action localization.

# 2 Local feature representation of video

Visual information of a video $\mathcal{V}$ is defined by a collection of its pixels $I$, that is $\mathcal{V} \supset I(r, c, t, \iota)$ with coordinates $(r, c, t)$ *(row, column, time)* and intensity $\iota$. We approach video action in an analogous way, decomposing an action $\mathcal{A}$ into local salient patches $x$, extracted around interest points, and represented as a quantized histogram of shape and motion flow gradients.

## 2.1 Feature detection and descriptor

In our work, we use Space Time Interest Points (STIP) developed from Laptev Laptev [2005] to detect local spatio-temporal features for human action in video.

**Space time interest point** The main idea of STIP is to extend Harris interest point detector Harris and Stephens [1988] from 2D image to 3D video, trying to find the point which has significant changes in both directions of space and time Laptev [2005]. The interest points are detected by searching for the pixel with high gradient change in shape and motion. Interest point location is represented by the triplet $(r, c, t)$ and written in short as $(\cdot)$. A filter constructed from a spatio-temporal second-moment matrix $\mu(r, c, t; \sigma_H, \tau_H)$ is used across all cube patches of the video, with

$$\mu(\cdot; \sigma_H, \tau_H) = g(\cdot; s\sigma_H, s\tau_H) * \Big( \bigtriangledown L(\cdot; \sigma_H, \tau_H) \Big( \bigtriangledown L(\cdot; \sigma_H, \tau_H) \Big)^T \Big), \quad (2.1)$$

where $\sigma_H$ represents Harris spatial detection scale, $\tau_H$ is on temporal direction. In Equation 2.1, $\bigtriangledown L$ is the space-time gradient function and $g$ is the separable Gaussian for smoothing purpose, which also applied at all video location

$$g(\cdot; \sigma_H^2, \tau_H^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma_H^4 \tau_H^4}} \times exp\Big( -(c^2 + r^2)/2\sigma_H^2 - t^2/2\tau_H^2 \Big). \quad (2.2)$$

Similar to spatial Harris corner and other interest point approach, STIP detects all local maxima of a threshold function $H^+ = (H > 0)$, with

$$H = det(\mu) - ktrace^3(\mu), \quad (2.3)$$

and selects those with positive values as space-time interest point. Figure 2.1 shows a few result snapshots of STIP extracted from TRECVid dataset at different scale levels.

**Histogram descriptor** At detected local feature $x$, a feature vector is extracted from the surrounding 3D video patch, having spatial-temporal size of $(\triangle_c(\sigma_H), \triangle_r(\sigma_H), \triangle_t(\tau_H))$, producing a feature vector $x = (r, c, t, s, z)$ with $s$ specifies its scale in region radius, $z$ is the feature description $z = (HoG, HoF)$, representing appearance and motion information at $x$ as Histogram of oriented Gradients (HoG) Dalal and Triggs [2005] and Histogram of oriented Flows (HoF) respectively. HoG and HoF are concatenated in $z$ according to the descriptor size by $\triangle_c(\sigma_H) = \triangle_r(\sigma_H) = 18\sigma_H$,$\triangle_t(\tau_H) = 8\tau_H$.

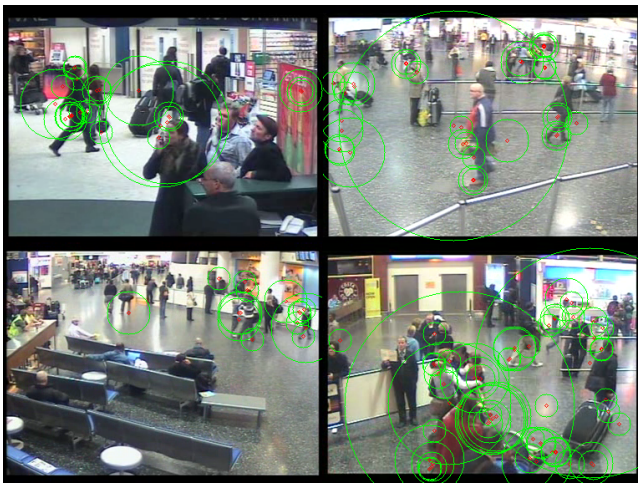


Figure 2.1: STIP detected at green circles on TRECVid.

## 2.2 Hierarchical Bayesian Feature Selection

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 Schuldt et al. [2004a] and Weizmann Blank et al. [2005] 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) Marszalek et al. [2009] 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. [2008] 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. [2008], we adopt a sparse kernel machine for classification purpose, with the function between the poste-

rior probability $p$ and probit link $\Phi$ defined in Tham et al. [2002]:

$$p(y_i = 1|x_i, \beta, \gamma) = \Phi\Big(q(x_i, \beta, \gamma)\Big), \tag{2.4}$$

with $q$ is the regression function

$$q(x_i, \beta, \gamma) = \sum_{k=1}^{N} \beta_k \gamma_k \psi(x_i, x_k), \tag{2.5}$$

and $\psi(x_i, x_k) = exp(-(x_i - x_k)/\sigma_R)$, the regression 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 Carbonetto et al. [2008].

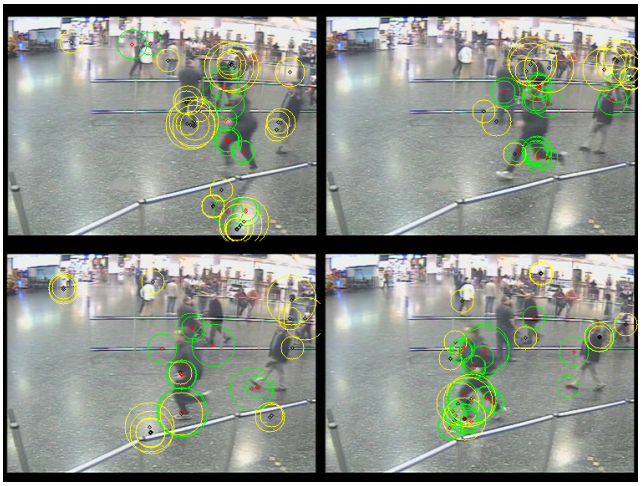


Figure 2.2: **HBFS** labels green circles as relevant features for action **PersonRuns**, yellow circles for noise which will be eliminated.

In order to increase the flexibility of the model, we adopt the idea described in Kuck et al. [2004] to assign both parameters $\beta$ and $\gamma$ with relevant distributions, respectively $\beta$ with an inverse Gamma distribution, and $\gamma$ with Beta distribution, hence comes the *Hierarchical* characteristic for this selection. The binary classification of label $y_i$ as shown in Carbonetto et al. [2008] is now the calibration of regression function $q(x_i, \beta, \gamma)$ (Equation 2.5) over zero.

$$y_i = \begin{cases} 1 & \text{if } q(x_i, \beta, \gamma) > 0 \\ -1 & \text{otherwise .} \end{cases} \tag{2.6}$$

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

$$p(y'|x', x, y_k) = \int p(y'|x', \theta) p(\theta|x, y^k) d\theta. \tag{2.7}$$

The computation of Equation 2.7 is clearly explained in Carbonetto et al. [2008] using Markov Chain Monte Carlo sampling in addition with a blocked Gibbs sampler as advised by Tham et al. [2002]. Figure 2.2 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 5.2.

# 3 Classification with Support Vector Machine

After the feature extraction task, each video shot can be seen as a sparse set of all feature points $x = (r, c, t, s, z, l)$ with label $l = 1$ indicating all these points belong to this action class of interest. We carry out action classification as a standard *bag-of-feature* approach, which consists of a quantization process of all selected local features, forming a histogram feature vector $h$ for each action candidate. In order to see effects of different kernel using local features, we put returned histograms into SVM with 2 types of kernel, namely *Linear* Cortes and Vapnik [1995]

$$K(h_i, h_j) = h_i^T h_j, \tag{3.1}$$

and $\chi^2$ Laptev et al. [2008b]

$$K(H_i, H_j) = exp\Big( - \frac{1}{A} D(H_i, H_j)\Big), \tag{3.2}$$

where $H_i = \{h_{in}\}$ and $H_j = \{h_{jn}\}$ are visual word histograms in $V$ dictionary and $D$ is $X^2$ distance function having training average $A$

$$D(H_i, H_j) = \frac{1}{2} \sum_{n=1}^{V} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}. \tag{3.3}$$

In this supervised learning task, we only use single feature channel, which is different from Laptev et al. [2008b], to show the actual improvement effect of HBFS on feature selection. The task of multi-class action classification is done using *one-against-all* approach, that is, when one action is used to build the classifier, all instances of other classes are considered as negative samples, and class label is assigned based on maximum prediction weights. Figure 3.1 shows the binary classification results of *ObjectPut* action classifier.


Figure 3.1: Using action model **ObjectPut**, SVM classifies the left video shot (blue text with (+) sign) as positive instance, and the right shot (red text with (-) sign) as negative.

# 4 Action localization with structured learning

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 Liu et al. [2009] or real world surveillance scenarios TRECVid Smeaton et al. [2006], 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 as *Integral Volume*, which basically is a bounding cuboid of all features $x|w(x) > \eta$, with $\eta$ 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 $\tau$ 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.

In order to see effects of different structured learning approach for the problem of action localization using a probabilistic *Dynamic Conditional Random Fields*, described in Section 4.1 and *max-margin Structured Support Vector Machines* in Section 4.2.

## 4.1 Dynamic Conditional Random Fields

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) Lafferty et al. [2001] are most appealing to our case of dependent sparse local features. For the task of object localization in images, Carbonetto et al. [2008] 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 Wang and Ji [2006] to develop a Dynamic Conditional Random Fields (DCRF) with an extra temporal constraint. Wang and Ji in Wang and Ji [2006] 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 feature x, the 3D cuboid

extracted around STIP, as the feature observations, shown as small green rectangles in Figure 4.2(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

$$p(w_k|z_{1:k}) \propto exp\{-\sum_{x\in X}[\varphi_x\Big(w_k(x)|z_{1:k}\Big)+ \sum_{y\in N_x}\varphi_{x,y}\Big(w_k(x),w_k(y)|z_{1:k}\Big)]\}. \tag{4.1}$$

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

$$p(w_{k+1}|w_k) \propto exp\{-\sum_{x\in X}[\varphi_x\Big(w_{k+1}(x)|w_k(N'_x)\Big)+ \sum_{y\in N_x}\varphi_{x,y}\Big(w_{k+1}(x),w_{k+1}(y)\Big)]\}, \tag{4.2}$$

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 Wang and Ji [2006] as

$$p(z_k|w_k) \propto exp\{-\sum_{x\in X}[\varphi_x\Big(z_k|w_k(x)\Big)+ \sum_{y\in N_x}\varphi_{x,y}\Big(z_k(x),z_k(y)|w_k(x),w_k(y)\Big)]\}, \tag{4.3}$$

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

$$\begin{aligned} p(z_k|w_k) &= p(hog_k, hof_k|w_k) \\ &= p(hog_k|w_k)p(hof_k|w_k). \end{aligned} \tag{4.4}$$

The optimization process is carried out similarly to the segmentation sampling described in Wang and Ji [2006], by approximating the mean field probability $q_x(w_k(x)|z_{1:k})$

$$p(w_k|z_{1:k}) \approx \prod_{x\in X} q_x\Big(w_k(x)|z_{1:k}\Big), \tag{4.5}$$

$$\eta_D = \widehat{w}_k(x) = \arg\max_e q_x\Big(w_k(x) = e|z_{1:k}\Big). \tag{4.6}$$

where $e$ is the initialization value, $q_x\Big(w_o(x) = e\Big)$, and is set to 0.5 for all feature x in our case. The calculated $\widehat{w}_k(x)$ is the final feature weight of all feature in the video shot, which will be passed through a thresholding weight filter $\eta_D$. The final *Integral Volume* is calculated as the as the approximate bounding rectangular cuboid that contains all those high weight features. Figure 4.1 illustrates the localization results using DCRF for an instance of action *Embrace* from TRECVid dataset.


(a) Extracted cuboids at STIP location. Note that all these cuboids already passed through **HBFS**

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

Figure 4.1: Feature relevancy weighting using **DCRF**.

## 4.2 Localization with structured SVM weighting

Structured SVM (SSVM) is first introduced in Tsochantaridis et al. [2004] to do inference on interdependent and structured outputs. In this section we will formulate the problem of action localization using the framework described from Tsochantaridis et al. [2004]. Denoting $w$ as the weighting on one feature $i$ compared to another $j$, we want to find the best possible weighting arrangement $w^*$ that maximize a performance measurement, in this case, $\tau$. Here we first review the Kendall's *ranking performance measure* $\tau$ from Litchfield Jr and Wilcoxon [1955] particularly for feature weight $x_w$. $\tau$ is defined as a quantitative entity for measuring disagreement of two weighting $w_i$ and $w_j$, with disagreement weighting $Q$ - number of different ordering pairs using each type of weighting

$$\kappa(w_a, w_b) = 1 - \frac{2Q}{\binom{m}{2}}. \tag{4.7}$$

If we have the optimal weight arrangement $w^*$, we need to define a cost function $f$ to minimize the loss function $-\kappa(r_{f(\mathcal{A})}, w*)$, where

$$\kappa_P(f) = \int \kappa(w_{f_\mathcal{A}}, w^*) dP(\mathcal{A}, w*). \tag{4.8}$$

In order to efficiently find solution to the optimization problem in 4.8, we inherit max-margin Support Vector Machine (SVM) approach for ranking from

Joachims [2002] to learn from the supervised video-weight $(\mathcal{A}, w^*)$ pairs, that is, to find the weighting function $f$ that optimizes the equivalent empirical $\kappa$

$$\kappa_S(f) = \frac{1}{n}\sum_{i=1}^{n} \kappa(w_{f(\mathcal{A}_i)}, w_i^*). \tag{4.9}$$

The weight estimation of two different features $z_i$ and $z_j$ can now be represented as a SVM inequality constraint incorporated in the weighting function $f_{\overrightarrow{\omega}}(\mathcal{A})$ of each action class $\mathcal{A}$

$$(z_i, z_j) \in f_{\overrightarrow{\omega}}(\mathcal{A}) \Longleftrightarrow \overrightarrow{\omega}\Upsilon(\mathcal{A}, x_i) > \overrightarrow{\omega}\Upsilon(\mathcal{A}, x_i), \tag{4.10}$$

where $\overrightarrow{\omega}$ is the weight vector representing the max-margin coefficients in SVM hyperplane separation Scholkopf et al. [1997], and the $\Upsilon(\mathcal{A}, x_i)$ is the feature function that maps the action class with their selected local features. In our case $\Upsilon$ is selected as the a collection of local feature descriptor $z$ and its visual word cluster cohesiveness score $\varrho$, defined as the posterior density of assigning to a particular cluster, $k$, using centralization Gaussian $(\mu_C, \sigma_C)$:

$$\underbrace{\varrho}_{\text{Cohesiveness}} \propto \underbrace{\frac{|N_k|}{N}}_{\text{Prior}} \underbrace{\frac{1}{\sigma_{Ck}} \exp\Big(-\frac{(v_k - \mu_{Ck})^2}{2 * \sigma_{Ck}^2}\Big)}_{\text{Centralization}}. \tag{4.11}$$

The cluster prior determines the likeliness of one particular feature assigning to cluster $k$, and total cluster element number $v_k$. The cluster centralization term decides on the likelihood of this feature in current cluster.

The local features are in fact the visual word clusters, we have a set of inequality constraints as followed

$$\forall(z_i, z_j) \in w_1^* : \overrightarrow{\omega}\Upsilon(\mathcal{A}_1, z_i) > \overrightarrow{\omega}\Upsilon(\mathcal{A}_1, z_j), \tag{4.12}$$

$$\forall(z_i, z_j) \in w_n^* : \overrightarrow{\omega}\Upsilon(\mathcal{A}_n, z_i) > \overrightarrow{\omega}\Upsilon(\mathcal{A}_n, z_j). \tag{4.13}$$

At this point, we now have the complete structured SVM optimization formulation defined as in Litchfield Jr and Wilcoxon [1955]

$$minimize : V(\overrightarrow{\omega}, \overrightarrow{\xi}) = \frac{1}{2}\overrightarrow{\omega} \cdot \overrightarrow{\omega} + C\sum \xi_{i,j,k}, \tag{4.14}$$

subject to:

$$\forall(z_i, zj) \in r_1^* : \overrightarrow{\omega}\Upsilon(\mathcal{A}_1, z_i) - \overrightarrow{\omega}\Upsilon(\mathcal{A}_1, z_j) \geq 1 - \xi_{i,j,1}, \tag{4.15}$$

$$\forall(z_i, zj) \in r_n^* : \overrightarrow{\omega}\Upsilon(\mathcal{A}_n, z_i) - \overrightarrow{\omega}\Upsilon(\mathcal{A}_n, z_j) \geq 1 - \xi_{i,j,n}, \tag{4.16}$$

$$\forall i \forall j \forall k : \xi_{i,j,k} \geq 0. \tag{4.17}$$

We implement this *max-margin* local feature weighting formulation based on the $SVM^{struct}$ framework from Tsochantaridis et al. [2004] to find the SVM margin weight configuration $\overrightarrow{\omega}^*$ via learning

$$(z_i, z_j) \in f_{\overrightarrow{\omega}}(\mathcal{A}), \tag{4.18}$$

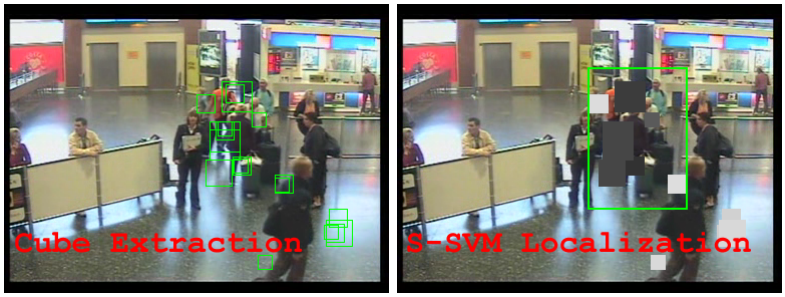
$$\Longleftrightarrow \overrightarrow{\omega}\Upsilon(\mathcal{A}, z_i) > \overrightarrow{\omega}\Upsilon(\mathcal{A}, z_j), \tag{4.19}$$

$$\Longleftrightarrow \sum \alpha^*_{k,l} \Upsilon(\mathcal{A}_k, z_l)\Upsilon(\mathcal{A}, z_i) > \sum \alpha^*_{k,l} \Upsilon(\mathcal{A}_k, z_l)\Upsilon(\mathcal{A}, z_j). \tag{4.20}$$

The final weighting for local features are then considered as the normalization of returned weighting in $(0, 1)$ which can be denoted as

$$\eta_S = \vec{\omega}\Upsilon(\mathcal{A}, z_i) = \sum \alpha^*_{k,l} \Upsilon(\mathcal{A}_k, z_l)\Upsilon(\mathcal{A}, z_j). \tag{4.21}$$

The final weighting results are then passed through a weighting filter $\eta_S$, which values vary across action class to form a final action volume boundary, which we call *Integral Volume*. We also use this weighting filter threshold to run different localization experiments and produce a Mean Average Precision, which will be reported in Section 5.3. Figure 4.2 shows a snapshot on *Embrace* action in TrecVID dataset using *structured SVM weighting*.




(a) Extracted cuboids at STIP location. Note that all these cuboids already passed through **HBFS**

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

Figure 4.2: Feature relevancy weighting using SSVM

# 5 Experimental results

## 5.1 Dataset selection and experiment setup

In order to evaluate performance of proposed approach, we run action classification and localization tasks on four datasets KTH Schuldt et al. [2004a], Weizmann Blank et al. [2005], Hollywood Human Action HOHA1 dataset Laptev et al. [2008a], and TRECVid 2008 Event Detection Development Set Smeaton et al. [2006].

**KTH** There are 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** There are about 90 colored video shots with 10 actions, *bend*, *jack*, *jump*, *pjump*, *run*, *side*, *skip*, *wave1*, *wave2*, *walk*, performed by 9 persons.

**HOHA** It contains 8 action classes, *AnswerPhone*, *GetOutCar*, *HandShake*, *HugPerson*, *Kiss*, *SitDown*, *SitUp*, and *StandUp*, distributed in around 450 training and testing videos of 448 manually annotated action labels.

**TRECVid** This is a challenging and realistic action dataset in surveillance video, recorded from 4 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, we extract all 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, *PersonRuns* 281 shots, and *Pointing* 1452 shots.

Figure 5.1 shows 8 detailed output stages of our action classification and localization framework which is used to evaluate effectiveness of structured learning in human action analysis. In addition, we summarize all main parameters used in our framework with their initial values in Table 5.1.

## 5.2 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 in Laptev et al. [2008a], which is 219 for training and 211 for testing, and lastly on TrecVID, we use *2/3 Split* for each action class. Performance on KTH and Weizmann is evaluated using *average accuracy* of classification confusion matrices, while on HOHA and TRECVid, we use *mean average precision* (MAP) to compare with reported works. For each dataset, we run cross combination of 2 local feature extraction, STIP without HBFS, and STIP with HBFS, associated with 2 SVM kernels, *Linear* and $\chi^2$.

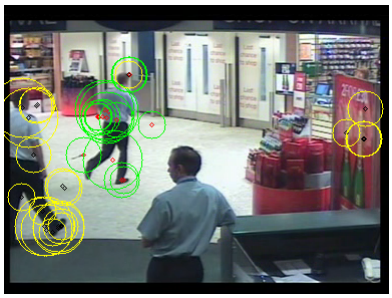
The results obtained from running the classifier on KTH and Weizmman are shown in confusion matrix Figure 5.2 and 5.3. The big improvement of classifier

| Symbol | Description | Eq. | Values |
|---|---|---|---|
| *Feature extraction* Section 2.1 | | | |
| $\sigma_H^2$ | Spatial Gaussian variance | 2.2 | *(4.0,8.0)* |
| $\tau_H^2$ | Temporal Gaussian variance | 2.2 | *(2.0,4.0)* |
| $k$ | Harris parameter | 2.3 | *5e-5* |
| $H$ | Detection threshold | 2.3 | *1e-12* |
| $\|z\|$ | HoG-HoF feature length | 4.1 | *162* |
| *Feature selection* Section 2.2 | | | |
| $\beta$ | initial $InvGamma$(shape-scale) | 2.4 | *(3.0-0.5)* |
| $\gamma$ | initial $Be$(shape-shape) | 2.4 | *(2.0-2.0)* |
| *Action classification* Section 3 | | | |
| $\|h\|$ | KTH visual dictionary size | 3.1 | *1024* |
| | Weizmann visual dictionary size | 3.1 | *64* |
| | HOHA visual dictionary size | 3.1 | *512* |
| | TRECVid visual dictionary size | 3.1 | *1024* |
| *Action localization DCRF* Section 4.1 | | | |
| $\|N\|$ | Neighborhood size (space-time) | 4.2 | *(5-3)* |
| $e$ | Initial mean field probability | 4.5 | *0.5* |
| $\eta_D$ | DCRF feature weight threshold | 4.6 | *0.25* |
| *Action localization SSVM* Section 4.2 | | | |
| $\varrho$ | Smoothing Gaussian cohesiveness | 4.11 | *(0,15)* |
| $\xi$ | Training error and margin trade-off | 4.14 | *1e-3* |
| $\eta_S$ | SSVM feature weight threshold | 4.21 | *0.5* |

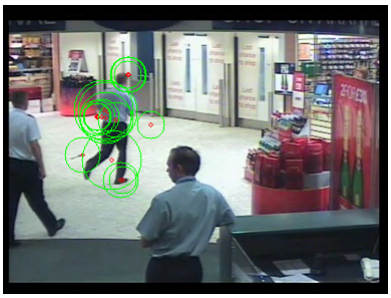
Table 5.1: Parameter summary and initialization

(a) Stage 1: Original frame

(b) Stage 2: STIP



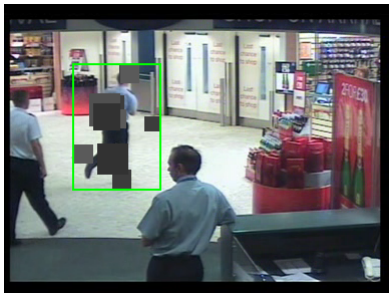
(c) Stage 3: HBFS



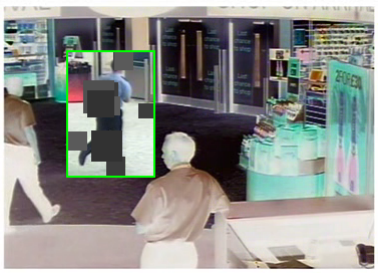
(d) Stage 4: SVM classification

(e) Stage 5: Extracted features

(f) Stage 6: Feature weighting

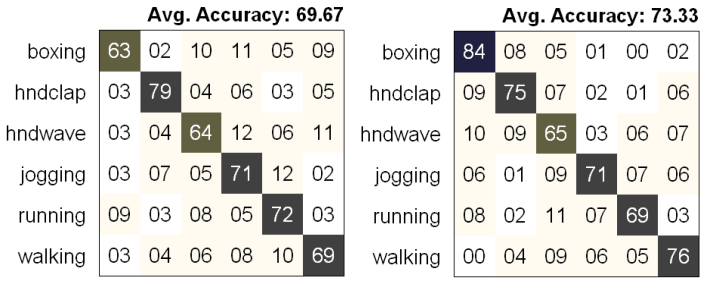


(g) Stage 7: Integral volume

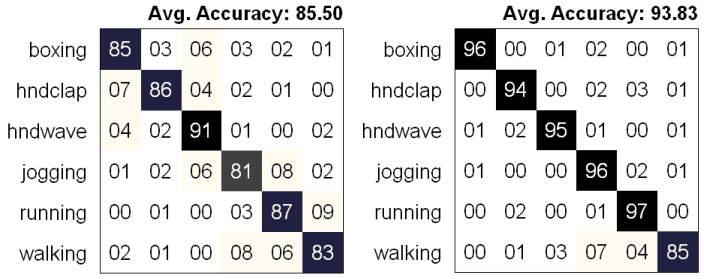


(h) Stage 8: Action localization

Figure 5.1: Detailed steps for recognizing action *PersonRuns* from TRECVid Event Detection Track

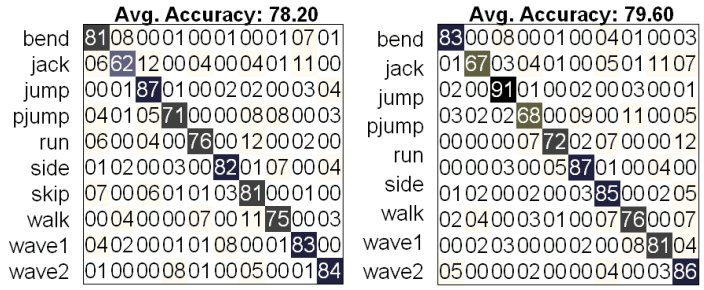



(a) Linear (b) $\chi^2$

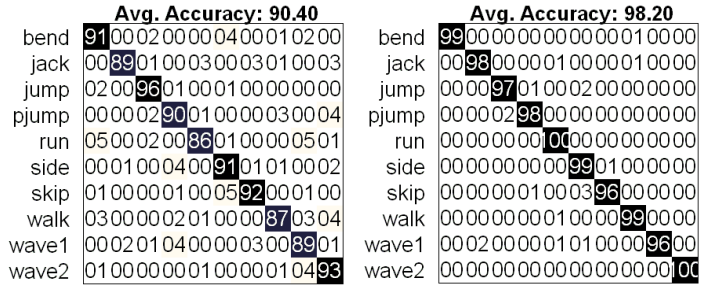



(c) HBFS - Linear (d) HBFS - $\chi^2$

Figure 5.2: Confusion matrix for action classification on KTH




(a) Linear (b) $\chi^2$




(c) HBFS - Linear (d) HBFS - $\chi^2$

Figure 5.3: Confusion matrix for action classification on Weizmann

using HBFS over non-HBFS has proved the effectiveness of our feature selection module. It can also be seen that $\chi^2$ kernel produces marginally better results than *Linear* kernel in most of the cases, and advantage of HBFS on $\chi^2$ is slightly better on *Linear*. We also compare our experiment results with those reported

| **Approach** | **KTH** | **Weizmann** |
|---|---|---|
| **HBFS-$\chi^2$** | ***93.8***% | ***98.2***% |
| **HBFS-Linear** | *85.5*% | *90.4*% |
| $\chi^2$ (baseline) | *73.3*% | *79.6*% |
| **Linear** (baseline) | *69.7*% | *78.2*% |
| Weinland and Boyer [2008] | * | **100**% |
| Gorelick et al. [2004] | * | 99.6% |
| Lin et al. [2009] | **95.8**% | * |
| Liu and Shah [2008] | 94.2% | * |
| Sun and Hauptmann [2009] | 94.0% | 97.8% |
| Grundmann et al. [2008] | 93.5% | 96.4% |
| Mikolajczyk and Uemura [2008] | 93.2% | * |
| Schindler and Van Gool [2008] | 92.7% | **100**% |
| Laptev et al. [2008b] | 91.8% | * |
| Jhuang et al. [2007] | 91.7% | 98.8% |
| Wang and Mori [2009] | 91.2% | 98.3% |
| Fathi and Mori [2008] | 90.5% | **100**% |
| Rapantzikos et al. [2009] | 88.3% | * |
| Jiang et al. [2006] | 84.4% | * |
| Willems et al. [2008] | 84.4% | * |
| Niebles et al. [2008b] | 81.5% | 72.8% |
| Dollar et al. [2005b] | 81.2% | * |
| Ke et al. [2007b] | 80.9% | * |
| Schuldt et al. [2004b] | 71.7% | * |

Table 5.2: Action classification performance comparison on KTH and Weizmann

in the literature in Table 5.2. It can be seen that even though we only use single channel SVM kernel, HBFS-$\chi^2$ (93.83%) still outperforms multi-channel Gaussian kernel of Laptev et al. [2008b] (91.80%). It is also worth mentioning that those approaches in Weinland and Boyer [2008], Lin et al. [2009], Liu and Shah [2008] and Sun and Hauptmann [2009] which give better results that ours are actually using holistic approach with a *pre-engineering* foreground motion extraction, which were previously mentioned in Section 1.1.

Classification results on HOHA is shown in Table 5.3 as action class based, using MAP to compare with *state-of-the-art* works. HBFS-$\chi^2$ outperforms all single channel features approach Raptis and Soatto [2010]; Matikainen et al. [2009]; Klaser et al. [2008], while appears to be highly competitive with other multi-channel approaches in Laptev et al. [2008b]; Yeffet and Wolf [2009]; Sun et al. [2009].

Table 5.4 summarizes classification results on TRECVid, also in action based. Classification performance in this dataset again shows the big advantage of applying HBFS for local feature, and also non-linear kernel is a good

| **Approach** | **AnsPh** | **OutCar** | **HndSh** | **HugPsn** | **Kiss** | **SitDwn** | **SitUp** | **StndUp** | **MAP** |
|---|---|---|---|---|---|---|---|---|---|
| **HBFS-$\chi^2$** | 28.9% | **52.7**% | 26.0% | 36.9% | 42.6% | **42.1**% | 18.7% | 33.9% | **35.2**% |
| **HBFS-Linear** | 24.4% | 30.8% | 24.7% | 36.2% | 47.6% | 39.7% | 17.7% | 38.1% | 32.4% |
| $\chi^2$ | 27.0% | 45.3% | 22.6% | 32.3% | 39.1% | 37.2% | 17.0% | 26.9% | 30.9% |
| **Linear** | 21.3% | 15.6% | 23.2% | 30.0% | 46.8% | 39.1% | 16.9 | 36.1% | 28.6% |
| Tracklet | | | | | | | | | |
| HoG-HoF BoF | 26.7% | 28.1% | 18.9% | 25.0% | 51.5% | 23.8% | **23.9**% | 59.1% | 32.1% |
| AoG-HoF BoF | 33.0% | 27.0% | 20.1% | 34.5% | 53.7% | 27.4% | 19.0% | **60.0**% | 34.3% |
| STIP | | | | | | | | | |
| Single | 26.7% | 22.5% | 23.7% | 34.9% | 52.0% | 37.8% | 15.2% | 45.4% | 32.9% |
| Combined | 32.1% | 41.5% | 32.3% | **40.6**% | 53.3% | 38.6% | 18.2% | 50.5% | 38.4% |
| Local | **35.1**% | 32.0% | **33.8**% | 28.3% | **57.6**% | 36.2% | 13.1% | 58.3% | 36.8% |
| Tracjectons | 35.0% | 7.7% | 5.3% | 23.5% | 42.9% | 13.6% | 11.1% | 42.9% | 22.8% |
| Spatio | 18.6% | 22.6% | 11.8% | 19.8% | 47.0% | 32.5% | 7.0% | 38.0% | 24.7% |
| Hierarchical | | | | | | | | | |
| TTD | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 30.3% |
| TTD-SIFT | | | | | | | | | **44.9**% |

Table 5.3: Mean Average Precision of action classification on HOHA

selection for local spatio-temporal features. In addition, it is quite clear that overall performance on KTH and Weizmann is larger better than on HOHA and TRECVid, which is quite reasonable due to the scenario complexity difference in these datasets.

| **Action** | **Linear** | $\chi^2$ | **HBFS-L_r** | **HBFS-$\chi^2$** |
|---|---|---|---|---|
| **CellToEar** | 26.6% | 31.3% | 34.7% | **36.4**% |
| **Embrace** | 30.8% | 19.0% | **36.7**% | 30.8% |
| **ObjectPut** | 19.7% | 24.9% | 23.6% | **27.0**% |
| **OpposingFlow** | 19.0% | 21.1% | **29.0**% | 22.1% |
| **PeopleMeet** | **21.2**% | 16.2% | 15.6% | 18.4% |
| **PeopleSplitUp** | **22.7**% | 15.2% | 21.2% | 20.7% |
| **PersonRuns** | 37.4% | 39.0% | 44.9% | **54.1**% |
| **Pointing** | 31.0% | 35.9% | 41.9% | **43.2**% |
| **MAP** | 26.1% | 25.3% | 31.0% | **31.6**% |

Table 5.4: Mean Average Precision of action classification on TRECVid

### 5.3 Action localization

As to our knowledge, we are the first who carry out localization performance on all 4 datasets KTH, Weizmann, HOHA, and TRECVid. Mean Average Precision is used on all datasets with weight threshold as ranking criteria, which was previously mentioned in Section 4.2. We train both DCRF and SSVM on all same training amount as used in action classification of Section 5.2. Feature mappings are initialized with 0 and 1 using ground-truth information of training instances. The predicted weights are then normalized and action volume are extracted at different weight thresholds. The overlapping between *ground-truth* data and estimated data of more than 50% is required for a action localization instance to be counted as true positive. Table 5.5 summarizes the localization results using *action-based* on 4 datasets.

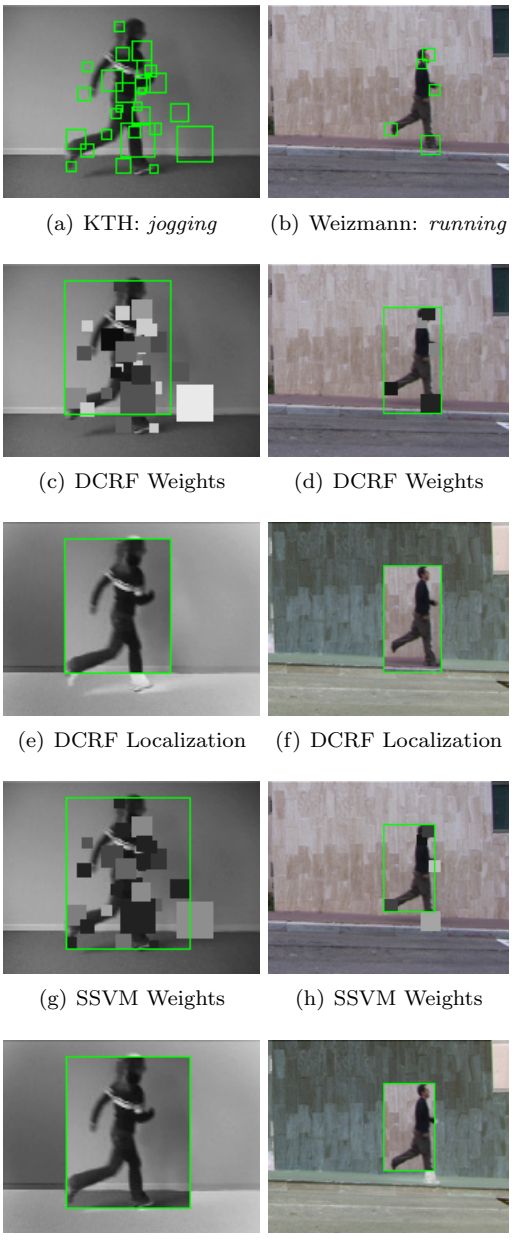
There are two main observation of action localization experiment. Firstly, on KTH and Weizmann, due to their *single-actor*, *uniform-background* characteristics, action localization is straightforward and only yield insignificant false localization. While on HOHA and TRECVid, localization is *non-trivial* with *cluttered-background* and *multi-actor* scenarios. Secondly, DCRF and SSVM yield slightly similar performance across all datasets, with minor difference in action types, which DCRF appears to work better with multi-actor activities, like *PeopleMeet*, *PeopleSplitUp*, *OpposingFlow*, where SSVM is more suitable for single actor localization, typically in *SitDown*, *Pointing*. Nevertheless, HBFS does prove to be also helpful for localization task, which on average improve around 7% in MAP across all datasets and actions. Figure 5.4 and 5.5 illustrates some localization results on the 4 datasets using two implementations of DCRF and SSVM.

## 6 Conclusion

We have presented a new framework for human action analysis by extensively utilizing the methods of structured learning. In particular, we formulate a feature selection step using a *hierarchical Bayesian machine* to filter sparse salient local features, which is shown to improve significantly over the existing *bag-of-feature* approaches. Secondly, we tackle the challenging task of action localization with two different structured learning approaches, one is *Dynamic Conditional Random Fields* based on probabilistic viewpoint, and the other is *structured Support Vector Machines* from max-margin principle. Empirical results on action testbeds demonstrate the potentials and applicability of our framework. For further work we would experiment with different action related datasets including interactive actions of multiple persons.

## 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.



(a) KTH: *jogging* (b) Weizmann: *running*

(c) DCRF Weights (d) DCRF Weights

(e) DCRF Localization (f) DCRF Localization

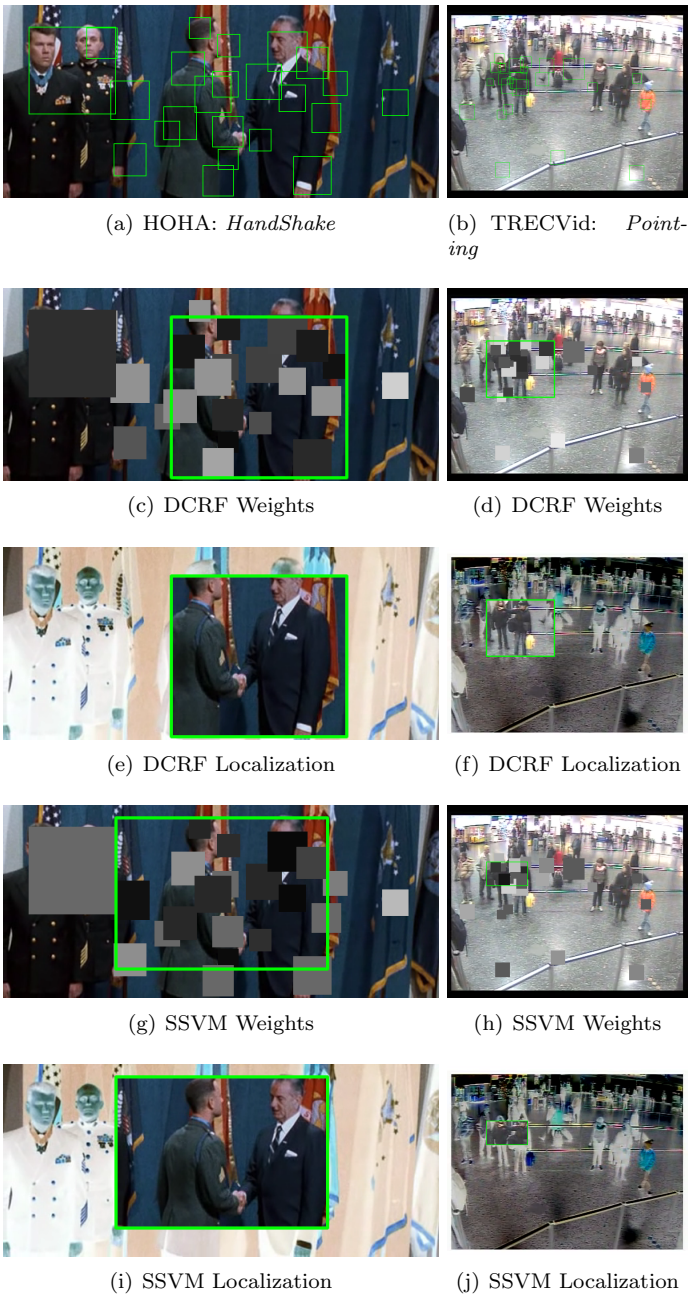
(g) SSVM Weights (h) SSVM Weights

(i) SSVM Localization (j) SSVM Localization

Figure 5.4: Sample keyframe snapshots for action localization results using DCRF and SSVM. Action of different datasets are shown on each row, including action *jogging* from KTH and *running* from Weizmann. First row shows the selected local features using HBFS, second and third row are weighting results obtained using DCRF, while the last two rows are results from SSVM.

# References

A. Bobick, J. Davis, The recognition of human movement using temporal templates, Pattern Analysis and Machine Intelligence, IEEE Transactions on



(a) HOHA: *HandShake* (b) TRECVid: *Pointing*

(c) DCRF Weights (d) DCRF Weights

(e) DCRF Localization (f) DCRF Localization

(g) SSVM Weights (h) SSVM Weights

(i) SSVM Localization (j) SSVM Localization

Figure 5.5: Sample keyframe snapshots for action localization results using DCRF and SSVM. Action of different datasets are shown on each row, including action *HandShake* from HOHA and *Pointing* from TRECVid. First row shows the selected local features using HBFS, second and third row are weighting results obtained using DCRF, while the last two rows are results from SSVM.

| Action | Without HBFS | | With HBFS | |
|---|---|---|---|---|
| | DCRF | SSVM | DCRF | SSVM |
| **KTH** | 84.3% | 83.4% | **95.1**% | 94.0% |
| **boxing** | 80.9% | 86.9% | **97.6**% | 96.6% |
| **handclapping** | 84.4% | 84.1% | 99.2%% | **100**% |
| **handwaving** | 90.8% | 86.4% | **100**% | 99.8% |
| **jogging** | 87.0% | 79.3% | **93.2**% | 90.7% |
| **running** | 80.6% | 81.1% | **87.9**% | 86.8% |
| **walking** | 82.1% | 82.6% | **92.9**% | 90.1% |
| | | | | |
| **Weizmann** | 98.0% | 97.8% | 98.7% | **99.2**% |
| **bend** | **100**% | **100**% | **100**% | **100**% |
| **jack** | **100**% | 99.3% | **100**% | **100**% |
| **jump** | 97.2% | **100**% | 99.3% | **100**% |
| **pjump** | 98.9% | **100**% | **100**% | **100**% |
| **run** | 89.6% | 89.2% | 92.2% | **95.3**% |
| **side** | **100**% | 93.3% | **100**% | 99.1% |
| **skip** | 98.0% | 99.0% | 97.8% | **100**% |
| **walk** | 97.4% | 97.5% | **98.1**% | 98.0% |
| **wave1** | **100**% | **100**% | **100**% | **100**% |
| **wave2** | **100**% | **100**% | **100**% | **100**% |
| | | | | |
| **HOHA** | 67.6% | 66.8% | 73.6% | **73.4**% |
| **AnswerPhone** | 61.0% | 61.4% | 65.0% | **65.6**% |
| **GetOutCar** | 71.9% | 77.8% | 88.9 | **90.1**% |
| **HandShake** | 62.8% | 62.2% | **67.2**% | 63.9% |
| **HugPerson** | 69.3% | 60.8% | **71.4**% | 61.8% |
| **Kiss** | 75.2% | 78.0% | 83.9% | **92.1**% |
| **SitDown** | 72.3% | 71.0% | 81.2% | **87.6**% |
| **SitUp** | 63.4% | 60.7% | **64.3**% | 63.6% |
| **StandUp** | 65.2% | 62.1% | **66.9**% | 62.8% |
| | | | | |
| **TRECVid** | 66.2% | 64.2% | **72.0**% | 70.4% |
| **CellToEar** | 58.9% | 61.1% | **72.3**% | 71.2% |
| **Embrace** | 70.8% | 70.8% | 71.3% | **89.1**% |
| **ObjectPut** | 61.0% | 59.7% | **69.0**% | 62.0% |
| **OpposingFlow** | 61.9% | 58.8% | **62.1**% | 60.6% |
| **PeopleMeet** | 71.0% | 74.8% | **84.3**% | 77.8% |
| **PeopleSplitUp** | 69.4% | 62.4% | **75.7**% | 68.4% |
| **PersonRuns** | 62.1% | 61.0% | 62.0% | **63.8**% |
| **Pointing** | 74.7% | 64.7% | **79.1**% | 70.0% |

Table 5.5: Mean Average Precision of action localization results.

23 (3) (2001) 257–267, ISSN 0162-8828.

D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using

motion history volumes, Computer Vision and Image Understanding 104 (2-3) (2006) 249–257.

A. Efros, A. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: IEEE International Conference on Computer Vision, vol. 2, Citeseer, 726–733, 2003.

A. Yilmaz, M. Shah, Actions Sketch: A Novel Action Representation, IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1 (2005) 984–989, ISSN 1063-6919.

L. Zelnik-Manor, M. Irani, Event-Based Analysis of Video, IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2 (2001) 123, ISSN 1063-6919.

Y. Ke, R. Sukthankar, M. Hebert, Event Detection in Crowded Videos, in: Proc. Intl. Conf. Computer Vision, 2007a.

L. Sigal, S. Bhatia, S. Roth, M. Black, M. Isard, Tracking loose-limbed people, Proc. IEEE Conf. Computer Vision and Pattern Recognition ISSN 1063-6919.

D. Ramanan, D. Forsyth, A. Zisserman, Tracking People by Learning Their Appearance, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1) (2007) 65–81.

H. Moon, R. Chellappa, 3D shape-encoded particle filter for object tracking and its application to human body tracking, Journal on Image and Video Processing 2008 (2008) 1–16, ISSN 1687-5176.

M. Abdelkader, A. Roy-Chowdhury, R. Chellappa, U. Akdemir, Activity representation using 3D shape models, Journal on Image and Video Processing 2008 (2008) 1–16, ISSN 1687-5176.

B. Stenger, A. Thayananthan, P. Torr, R. Cipolla, Model-based hand tracking using a hierarchical bayesian filter, IEEE Transactions on Pattern Analysis and Machine Intelligence (2006) 1372–1384ISSN 0162-8828.

F. Guo, G. Qian, Monocular 3D tracking of articulated human motion in silhouette and pose manifolds, Journal on Image and Video Processing 2008 (2008) 4, ISSN 1687-5176.

A. Oikonomopoulos, I. Patras, M. Pantic, Spatiotemporal salient points for visual recognition of human actions, IEEE Tracsactions on Systems, Man, and Cybernetics 36 (3) (2006) 710–719.

T. Kadir, M. Brady, Scale saliency: A novel approach to salient feature and scale selection, in: Visual Information Engineering, 2003. VIE 2003. International Conference on, IET, ISBN 0852967578, ISSN 0537-9989, 25–28, 2003.

A. Fathi, G. Mori, Action recognition by learning mid-level motion features, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.

M. Ahmad, S. Lee, Human action recognition using shape and CLG-motion flow from multi-view image sequences, Pattern Recognition 41 (7) (2008) 2237–2252, ISSN 0031-3203.

E. Shechtman, M. Irani, Space-time behavior based correlation, IEEE Transactions on Pattern Analysis and Machine Intelligence ISSN 1063-6919.

I. Laptev, On Space-Time Interest Points, Intl. Journal of Computer Vision (2-3) (2005) 107–123.

P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: PETS, 2005a.

C. Harris, M. Stephens, A combined corner and edge detector, in: Alvey vision conference, vol. 15, Manchester, UK, 50, 1988.

C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proc. Intl. Conf. Pattern Recognition, 2004a.

J. Niebles, H. Wang, F. Li, Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words, IJCV 79 (3).

T. Thi, L. Cheng, J. Zhang, L. Wang, Implicit Motion-Shape Model: A generic approach for action matching, in: Image Processing (ICIP), 2010 17th IEEE International Conference on, ISSN 1522-4880, 1477–1480, 2010a.

T. Thi, L. Cheng, J. Zhang, L. Wang, S. Satoh, Weakly Supervised Action Recognition using Implicit Shape Models, in: 2010 International Conference on Pattern Recognition, IEEE, ISSN 1051-4651, 3517–3520, 2010b.

M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as Space-Time Shapes, in: ICCV, 2005.

J. Yuan, Z. Liu, Y. Wu, Discriminative subvolume search for efficient action detection, Proc. IEEE Conf. Computer Vision and Pattern Recognition .

J. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, International Journal of Computer Vision 79 (3) (2008b) 299–318.

A. Oikonomopoulos, I. Patras, M. Pantic, Spatiotemporal Localization and Categorization of Human Actions in Unsegmented Image Sequences, Image Processing, IEEE Transactions on (99) (2010) 1–1, ISSN 1057-7149.

K. Alexander, Human Focused Action Localization in Video, Laboratory investigation a journal of technical methods and pathology 84 (5), ISSN 0023-6837.

I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: CVPR, 2008a.

A. F. Smeaton, P. Over, W. Kraai, Evaluation campaigns and TRECVid, in: MIR, 2006.

T. Thi, J. Zhang, L. Cheng, L. Wang, S. Satoh, Human Action Recognition and Localization in Video Using Structured Learning of Local Space-Time Features, in: Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on, IEEE, 204–211, 2010c.

B. Scholkopf, A. Smola, K. Muller, Kernel principal component analysis, Proc. Intl. Conf. Artificial Neural Networks (1997) 583–588.

N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, Citeseer, 886, 2005.

M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

P. Carbonetto, G. Dorko, C. Schmid, H. Kuck, N. de Freitas, Learning to Recognize Objects with Little Supervision, Intl. Journal of Computer Vision 77 (1-3) (2008) 219–237.

S.-S. Tham, A. Doucet, K. Ramamohanarao, Sparse Bayesian Learning for Regression and Classification using Markov Chain Monte Carlo, in: Proc. Intl. Conf. Machine Learning, 2002.

H. Kuck, P. Carbonetto, N. de Freitas, A Constrained Semi-supervised Learning Approach to Data Association, 2004.

C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297, ISSN 0885-6125.

I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, Proc. IEEE Conf. Computer Vision and Pattern Recognition 1 (2008b) 20–23.

J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos 'in the wild', in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proc. Intl. Conf. Machine Learning, 2001.

Y. Wang, Q. Ji, A dynamic conditional random field model for object segmentation in image sequences, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006.

I. Tsochantaridis, T. Hofmann, T. Joachims, Y. Altun, Support vector machine learning for interdependent and structured output spaces, in: Proceedings of the twenty-first international conference on Machine learning, ACM, 104, 2004.

J. Litchfield Jr, F. Wilcoxon, Rank Correlation Method, Analytical Chemistry 27 (2) (1955) 299–300, ISSN 0003-2700.

T. Joachims, Optimizing search engines using clickthrough data, in: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, ISBN 158113567X, 133–142, 2002.

D. Weinland, E. Boyer, Action recognition using exemplar-based embedding, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 13, 2008.

L. Gorelick, M. Galun, E. Sharon, R. Basri, A. Brandt, Shape representation and classification using the poisson equation, vol. 2, IEEE Computer Society; 1999, 2004.

Z. Lin, Z. Jiang, L. Davis, Recognizing actions by shape-motion prototype trees, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, ISSN 1550-5499, 444–451, 2009.

J. Liu, M. Shah, Learning human actions via information maximization, in: IEEE Computer Society Conference on Computer Vision and Pattern Recongition, 2008.

X. Sun, M. Hauptmann, Action recognition via local descriptors and holistic features, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

M. Grundmann, F. Meier, I. Essa, 3D shape context and distance transform for action recognition, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.

K. Mikolajczyk, H. Uemura, Action recognition with motion-appearance vocabulary forest, in: Proceedings, IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2008.

K. Schindler, L. Van Gool, Action snippets: How many frames does human action recognition require, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.

H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action recognition, in: Proc. Intl. Conf. Computer Vision, 2007.

Y. Wang, G. Mori, Human action recognition by semi-latent topic models, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (10) (2009) 1762–1774.

K. Rapantzikos, Y. Avrithis, S. Kollias, Dense saliency-based spatiotemporal feature points for action recognition, in: Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

H. Jiang, M. Drew, Z. Li, Successive convex matching for action detection, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006.

G. Willems, T. Tuytelaars, L. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector 23 (2008) 650–653.

P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, ICCV VS-PETS .

Y. Ke, R. Sukthankar, M. Hebert, Event detection in crowded videos, in: ICCV, 2007b.

C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 3, 2004b.

M. Raptis, S. Soatto, Tracklet descriptors for action modeling and video analysis, Computer Vision–ECCV 2010 (2010) 577–590.

P. Matikainen, M. Hebert, R. Sukthankar, Trajectons: Action recognition through the motion analysis of tracked features, in: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, IEEE, 514–521, 2009.

A. Klaser, M. Marszalek, C. Schmid, A spatio-temporal descriptor based on 3D-gradients, in: British Machine Vision Conference, Citeseer, 995–1004, 2008.

L. Yeffet, L. Wolf, Local trinary patterns for human action recognition, in: IEEE 12th International Conference on Computer Vision, IEEE, ISSN 1550-5499, 492–497, 2009.

J. Sun, X. Wu, S. Yan, L. Cheong, T. Chua, J. Li, Hierarchical spatio-temporal context modeling for action recognition, Proc. IEEE Conf. Computer Vision and Pattern Recognition .