Detection of Activities and Events without Explicit Categorization

Masakazu Matsugu [†] Masao Yamanaka [†] Masashi Sugiyama [‡]

[†] CANON Inc. Visual Information Technology Development Center, Corporate R&D Headquarters
[‡] Graduate School of Information Science and Engineering, Tokyo Institute of Technology
{matsugu.masakazu,yamanaka.masao}@canon.co.jp sugi@cs.titech.ac.jp

Abstract

We address the problem of unsupervised detection of events (e.g., changes or meaningful states of human activities) without any similarity test against specific models or probability density estimation (e.g., specific category learning). Rather than estimating probability densities, very difficult to calculate in general settings, we formulate the event detection as binary classification with density ratio estimation [9] in a hierarchical probabilistic framework. The proposed method takes pairs of video stream data (i.e., past and current) as input with differing time-scales, generates density ratio models in a way of online learning, and judges if there is any 'meaningful difference' between them based on the multiple density ratio estimations. Through experimental studies on real-world scenes of specific domains using challenging datasets from sports scene (i.e., tennis match) with complex background, we demonstrate the potential advantage of our approach over the state-of-the-art in terms of precision and efficiency.

1. Introduction

Analysis of events and human actions is very important because of various applications such as content-based video retrieval, visual surveillance, and human-computer interaction. Event detection and human action recognition are both challenging computer vision problems [1, 14]. A source of difficulty in recognizing events arises from complexity and variations of backgrounds, contexts [26] and events themselves as well. Variation of clothes, sizes, or postures of people, illumination conditions, occlusion conditions, and camera angels are other factors that render the problem more challenging.

The task of automatic semantic categorization of events or actions generally demands huge amount of video data with ground truth annotations or segmentation of streamlined data [7]. In recent years, unsupervised approaches have been proposed for learning human action categories [12], [16] and also for detecting abnormal behavior [20].

Most action (event) detection methods extract features such as optical flow based features [12, 20], spatio-temporal features [5, 10-15, 27], or static features including appearance, shape, and spatial relations of local features [8, 18]. Some unsupervised approaches, after extraction of those primitive features, utilize codebook representation [2, 13, 19] effective for describing and discriminating various event categories. BoW representation of spatio-temporal interest points in particular has received considerable attention for the recognition of actions and events as well [2, 13, 21, 26]. Technologies for detecting semantic events can be summarized such as finite state machines, statistical models (e.g., HMM [6, 17], Bayesian networks [3]), kernel methods [4, 18, 21], and tree-structured classifiers [12].

In this paper, we define 'events' as changes or meaningful states of human activities that can be observed as visual contents, something significantly different from 'normal' states which are learned or statistically described by using past sequences of video data. We address the problem of unsupervised detection of semantic event categories by using density ratio in a hierarchical framework with differing time-scales. In general, estimation of accurate probability densities is known to be very difficult, and the proposed method dispenses with direct estimation of them (or any models, including matching kernels and templates), which correspond to semantic event category.

Main contribution of this paper is twofold. 1) the possibility of detecting semantic events by direct estimation of probability density ratio which is more accurate than estimating the probability densities of corresponding category followed with taking their ratio; 2) a new scheme for unsupervised learning of semantic events which requires much less sequence data than existing technologies, indicating the possibility of on-line learning.

Using the density ratio estimation [9], obtained from pairs of sequence data described by spatio-temporal features (i.e. CHLAC [10]), we compute measures of 'anomaly' with differing time-scales, some of which possibly reflects semantic action category, a meaningful change of actions based on observed, past sequence data as reference. Thus, the proposed algorithm does not involve any probabilistic generative model, nor does require any sophisticated codebook representation as in [2, 13]. In the

previous studies, some approach proposed to reduce the semantic gap [22]. We do not claim that the proposed approach guarantees to detect any semantic events. However, we claim that the proposed hierarchical system can detect significant changes of visual contents as potential events by taking appropriate length of 'past' and 'current' sequential data, since an event can be characterized by unique time-scale (e.g., duration defined by the beginning and end of event is equivalent to the time-scale).

The advantage of the proposed scheme is that, in principle, it can deal with any complex probability density related with event category, since only density ratio does matter in the algorithm. Thus, in the training stage, we do NOT assume a single person performing only one action [13].

The remainder of this paper is organized as follows. At first, we formulate the video event detection problem based on the density-ratio estimation in Section 2. In Section 3, we report experimental results on real-world video sequences. Finally, we conclude by summarizing our contribution in Section 4.

2. Problem formulation and proposed approach

In this section, we formulate a video event detection problem based on the probability density ratio.

2.1. Problem formulation

Let $\mathbf{x}(t)$ be a d-dimensional feature vector at time t. Our task is to detect whether there exists a change point based on video event between two consecutive time intervals, respectively called the *train* and *test* intervals.

A naive approach to this problem would be to first estimate the *train* and *test* probability density functions separately from the *train* and *test* time series features, and then calculate the difference between *train* and *test* intervals by taking the estimated probability density functions.

However, since non-parametric density estimation is known to be a hard problem [23], this naive approach to change point detection via non-parametric density estimation may not be effective; directly estimating the probability density ratio without estimating the probability densities would be more promising.

Thus, our algorithm is based on the probability density ratio of the time series features $\mathbf{x}(t)$ defined by

$$w(\mathbf{x}) = \frac{p_{te}(\mathbf{x})}{p_{tr}(\mathbf{x})},$$

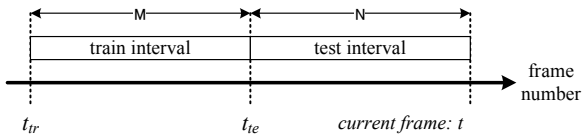

Figure 1: Definition of train and test intervals

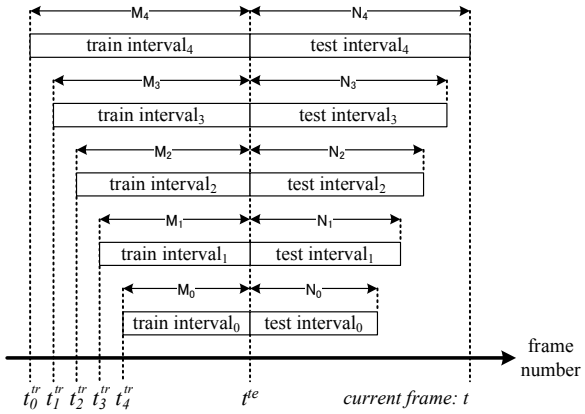

Figure 2: Definition of train and test intervals in a hierarchy

where $p_{te}(\mathbf{x})$ and $p_{tr}(\mathbf{x})$ are the probability density functions of the *test* and *train* time series features, respectively.

Therefore, we can decide whether there is a change point based on video event between the *train* and *test* intervals by monitoring the logarithm of the probability density ratio:

$$S = \sum_{t=t_{tr}}^{t_{te}} \ln\left(\frac{p_{te}(\mathbf{x})}{p_{tr}(\mathbf{x})}\right),$$

where t_{tr} and t_{te} $(t_{tr} < t_{te})$ are respectively the starting time points of the *train* and *test* intervals, The score of S serves as a measure of 'anomaly' assuming that the *train* data indicating 'normal' state, and it is also a measure of event detection. The above formulation is summarized in Figure 1, where M and N are respectively the number of frame in the train and test intervals.

Based on the value of S, we can conclude that

$$\begin{cases} S \le \mu \;\rightarrow\; \textit{no video event occurs}, \\ \textit{otherwise} \;\rightarrow\; \textit{a video event occurs}, \end{cases}$$

where $\mu(>0)$ is a predetermined threshold for the detection of events.

But in practice, it is difficult to determine the *train* and *test* intervals to properly detect video events without any preconditions for every video sequences, so we should setup several intervals hierarchically in Figure 2. This hierarchical structure makes it possible to detect a variety of events from micro to macro level. Here, the score S_h of the h-th hierarchy is defined by

$$S_h = \sum_{t=t_h^{tr}}^{t^{te}} \ln\left(\frac{p_h^{te}(\mathbf{x})}{p_h^{tr}(\mathbf{x})}\right),$$

where t_h^{tr} $(< t^{te})$ are the starting time points of the train interval, and $p_h^{te}(\mathbf{x})$ and $p_h^{tr}(\mathbf{x})$ are the probability density functions of the *test* and *train* time series features in the *h*-th hierarchy in Figure 2, respectively. Finally, we can obtain abnormal score *S* defined by

$$S = \max_h(S_h).$$

The remaining question of this procedure is how to calculate the time series features $\mathbf{x}(t)$ from video sequences and density ratios:

$$w_h(\mathbf{x}) = \frac{p_h^{te}(\mathbf{x})}{p_h^{tr}(\mathbf{x})}.$$

2.2. Cubic higher-order local auto-correlation

Spatio-temporal features as exemplified like spatio-temporal interest points and spatio-temporal shapes have received great attentions and been successfully used for detection of events and actions. We adopt another spatio-temporal features, Cubic Higher order Local Auto Correlation (CHLAC) [10]. Both the CHLAC and its extension [25] have been successfully used in action recognition, and we consider the method [28] based on the CHLAC as one of the state-of-the-arts. This feature was chosen because of its three properties [10]: *additivity, shift invariance,* and *robustness to noise*. The CHLAC directly deals with *three* - dimensional data, suitable for motion image sequence.

Let $f(\mathbf{r})$ be three-way data with $\mathbf{r} = (x, y, z)$, the *N*-th order auto-correlation function is defined as

$$\mathbf{x}_N(\mathbf{a}_1,...,\mathbf{a}_N) = \int f(\mathbf{r}) f(\mathbf{r}+\mathbf{a}_1)\cdots f(\mathbf{r}+\mathbf{a}_N) d\mathbf{r} \quad (1)$$

where $\mathbf{a}_i (i = 1,\ldots,N)$ are displacement vectors to obtain positions to correlate with $f(\mathbf{r})$. In Eq.(1), displacement vector $\mathbf{a}_i$ is limited to a local region $3\times3\times3$ around $\mathbf{r}$ and the number of displacement vector *N* is set less than or equal to 2.

The CHLAC feature, the value of $\mathbf{x}_N(\mathbf{a}_1,...,\mathbf{a}_N)$, remains the same if the patterns of $(\mathbf{r},\mathbf{a}_1,\ldots,\mathbf{a}_N)$ are identical in the point configuration (i.e., *shift invariance* property). In the CHLAC features, such duplicate sets are eliminated, and example of mask patterns are shown in Figure 2 which indicate $\mathbf{r}$ and $\mathbf{a}_i$.

By taking inter-frame difference and thresholding, we convert input image data into motion-image sequences composed of binary data, for which 251 mask patterns were used (i.e., CHLAC is the 251-dimensional vector).

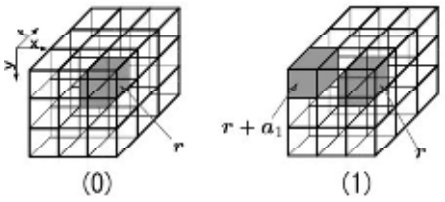

Figure 3: Example of a mask pattern [10]. (0) N=0; (1) N=1, a_1=(-1,-1,-1)'. Overlapping mask patterns are eliminated if mutually shifted in three-way data.

2.3. Direct density ratio estimation

As described in section 2.1, we need to estimate the density ratio for the event detection problem. Here, we show how the density ratio could be directly estimated without going through density estimation based on Unconstrained Least-Squares Importance Fitting (uLSIF) [9].

2.3.1 Formulation of Density Ratio Estimation

Suppose we are given a set of samples in the *h*-th hierarchy

$$\chi_h^{te} := \left\{\mathbf{x}_i^{te} \middle| \mathbf{x}_i^{te} \in R^d\right\}_{i=1}^{N_h}$$

drawn independently from a probability distribution P_h^{te} with density $p_h^{te}(\mathbf{x})$, and another set of samples in the *h*-th hierarchy

$$\chi_h^{tr} := \left\{\mathbf{x}_j^{tr} \middle| \mathbf{x}_j^{tr} \in R^d\right\}_{j=1}^{M_h}$$

drawn independently from (possibly) another probability distribution P_h^{tr} with density $p_h^{tr}(\mathbf{x})$:

$$\left\{\mathbf{x}_i^{te}\right\}_{i=1}^{N_h} \overset{i.i.d}{\sim} P_h^{te},$$

$$\left\{\mathbf{x}_j^{tr}\right\}_{j=1}^{M_h} \overset{i.i.d}{\sim} P_h^{tr}.$$

where, N_h and M_h are the number of samples of the *test* and *train* intervals in the *h*-th hierarchy, respectively.

The goal of density ratio estimation is to estimate the density ratio function

$$w_h(\mathbf{x}) := \frac{p_h^{te}(\mathbf{x})}{p_h^{tr}(\mathbf{x})}$$

from the samples χ_h^{te} and χ_h^{tr}, where we assume $p_h^{tr}(\mathbf{x}) > 0$ for all $\mathbf{x}$.

2.3.2 Least-Squares Approach to Density Ratio Estimation

Let us model the density ratio function $w_h(\mathbf{x})$ by the following kernel model:

$$\hat{w}_h(\mathbf{x}) := \sum_{i=1}^{N_h} \boldsymbol{\alpha}_i K(\mathbf{x}, \mathbf{x}_i) = \boldsymbol{\alpha}' \cdot \mathbf{k}(\mathbf{x}),$$

where

$$\boldsymbol{\alpha} := (\alpha_1, \alpha_2 ..., \alpha_{N_h})'$$

are parameters to be learned from data samples, $\bullet'$ denotes the transpose of a matrix or a vector,

$$\mathbf{k}(\mathbf{x}) := (K(\mathbf{x}, \mathbf{x}_1), K(\mathbf{x}, \mathbf{x}_2) ..., K(\mathbf{x}, \mathbf{x}_{N_h}))'$$

are kernel basis functions. A popular choice of the kernel is the Gaussian function:

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2} \right), \quad (2)$$

where σ^2 denotes the Gaussian variance.

We determine the parameter $\boldsymbol{\alpha}$ in the model $\hat{w}_h(\mathbf{x})$ so that the following squared-error J_0 is minimized:

$$J_0 := \frac{1}{2} \int \left(\hat{w}_h(\mathbf{x}) - w_h(\mathbf{x}) \right)^2 p_h^{tr}(\mathbf{x}) d\mathbf{x}$$

$$= \frac{1}{2} \int \hat{w}_h(\mathbf{x})^2 p_h^{tr}(\mathbf{x}) d\mathbf{x} - \int \hat{w}_h(\mathbf{x}) p_h^{te}(\mathbf{x}) d\mathbf{x} + \frac{1}{2} \int w_h(\mathbf{x}) p_h^{te}(\mathbf{x}) d\mathbf{x},$$

where the last term is a constant and therefore can be safely ignored. Let us denote the first two terms by J:

$$J(\boldsymbol{\alpha}) := \frac{1}{2} \int \hat{w}_h(\mathbf{x})^2 p_h^{tr}(\mathbf{x}) d\mathbf{x} - \int \hat{w}_h(\mathbf{x}) p_h^{te}(\mathbf{x}) d\mathbf{x} \quad (3)$$

$$= \frac{1}{2} \boldsymbol{\alpha}' \mathbf{H} \boldsymbol{\alpha} - \mathbf{h}' \boldsymbol{\alpha},$$

where $\mathbf{H}$ is the $N_h \times N_h$ matrix defined by

$$\mathbf{H} := \int \mathbf{k}(\mathbf{x}) \mathbf{k}(\mathbf{x})' p_h^{tr}(\mathbf{x}) d\mathbf{x},$$

and $\mathbf{h}$ is the N_h - dimensional vector defined by

$$\mathbf{h} := \int \mathbf{k}(\mathbf{x}) p_h^{te}(\mathbf{x}) d\mathbf{x}.$$

2.3.3 Empirical Approximation

Since J contains the expectation over unknown densities $p_h^{te}(\mathbf{x})$ and $p_h^{tr}(\mathbf{x})$, we approximate the expectations by empirical averages. Then we obtain

$$\hat{J}(\boldsymbol{\alpha}) := \frac{1}{2M_h} \sum_{j=1}^{M_h} \hat{w}_h(\mathbf{x}^{tr}_j)^2 - \frac{1}{N_h} \sum_{i=1}^{N_h} \hat{w}_h(\mathbf{x}^{te}_i)$$

$$= \frac{1}{2} \boldsymbol{\alpha}' \hat{\mathbf{H}} \boldsymbol{\alpha} - \boldsymbol{\alpha}' \hat{\mathbf{h}},$$

where $\hat{\mathbf{H}}$ is the $N_h \times N_h$ matrix defined by

$$\hat{\mathbf{H}} := \frac{1}{M_h} \sum_{j=1}^{M_h} \mathbf{k}(\mathbf{x}^{tr}_j) \mathbf{k}(\mathbf{x}^{tr}_j)',$$

and $\hat{\mathbf{h}}$ is the N_h-dimensional vector defined by

$$\hat{\mathbf{h}} := \frac{1}{N_h} \sum_{i=1}^{N_h} \mathbf{k}(\mathbf{x}^{te}_i).$$

By including a regularization term, the uLSIF optimization problem is formulated as follows.

$$\hat{\boldsymbol{\alpha}} := \arg\min_{\boldsymbol{\alpha}} \left[\frac{1}{2} \boldsymbol{\alpha}' \hat{\mathbf{H}} \boldsymbol{\alpha} - \boldsymbol{\alpha}' \hat{\mathbf{h}} + \frac{\lambda}{2} \boldsymbol{\alpha}' \boldsymbol{\alpha} \right],$$

where $\boldsymbol{\alpha}' \boldsymbol{\alpha} / 2$ is a regularizer and $\lambda (\geq 0)$ is the regularization parameter that controls the strength of regularization. By taking the derivative of the above objective function with respect to the parameter $\boldsymbol{\alpha}$ and equating it to zero, we can analytically obtain the solution $\hat{\boldsymbol{\alpha}}$ as

$$\hat{\boldsymbol{\alpha}} = (\hat{\mathbf{H}} + \lambda \mathbf{I}_{N_h})^{-1} \hat{\mathbf{h}},$$

where $\mathbf{I}_{N_h}$ is the N_h-dimensional identity matrix. Finally, the density ratio estimator $\hat{w}_h(\mathbf{x})$ is given by

$$\hat{w}_h(\mathbf{x}) := \hat{\boldsymbol{\alpha}}' \mathbf{k}(\mathbf{x}).$$

Thus, the resulting score S_h in section 2.1 is given by

$$S_h = \sum_{t=t_h^{tr}}^{t^{te}} \ln(\hat{w}_h(\mathbf{x})).$$

Thanks to the analytic-form expression, uLSIF is computationally more efficient than alternative density ratio estimators which involve non-linear optimization.

2.3.4 Model Selection by Cross-Validation

The practical performance of uLSIF depends on the choice of the kernel function (e.g. the kernel width σ in the case of Gaussian kernel Eq.(2)) and the regularization parameter λ. Model selection of uLSIF is possible based on *cross-validation* with respect to the error criterion J defined by Eq.(3) [24].

More specifically, each of the sample sets $\chi_h^{te} = \{\mathbf{x}_i^{te}\}_{i=1}^{N_h}$ and $\chi_h^{tr} = \{\mathbf{x}_j^{tr}\}_{j=1}^{M_h}$ is divided into L disjoint sets $\{{}_l\chi_h^{te}\}_{l=1}^{L}$ and $\{{}_l\chi_h^{tr}\}_{l=1}^{L}$. Then an uLSIF solution $\hat{w}_l(\mathbf{x})$ is obtained using $\chi_h^{te} \backslash {}_l\chi_h^{te}$ and $\chi_h^{tr} \backslash {}_l\chi_h^{tr}$ (i.e., all samples without ${}_l\chi_h^{te}$ and ${}_l\chi_h^{te}$), and its J-value for the hold-out samples ${}_l\chi_h^{te}$ and ${}_l\chi_h^{tr}$ is computed as

$${}_l\hat{J}_h^{CV} := \frac{1}{2\left| {}_l\chi_h^{tr} \right|} \sum_{\mathbf{x}^{tr} \in {}_l\chi_h^{tr}} \hat{w}_l(\mathbf{x}^{tr})^2 - \frac{1}{\left| {}_l\chi_h^{te} \right|} \sum_{\mathbf{x}^{te} \in {}_l\chi_h^{te}} \hat{w}_l(\mathbf{x}^{te}),$$

where $|\chi|$ denotes the number of elements in the set χ. This procedure is repeated for l = 1,...,L, and the average of ${}_l\hat{J}_h^{CV}$ over all l is computed as

$$\hat{J}_h^{CV} := \frac{1}{L} \sum_{l=1}^{L} {}_l\hat{J}_h^{CV}.$$

Finally, the model (the kernel width σ and the regularization parameter λ in the current setup) that minimizes ${}_l\hat{J}_h^{CV}$ is chosen as the most suitable one.

3. Experiments

In this section, we show two experimental studies to evaluate the proposed method and compare the results with the reference method [10, 28] that also uses the CHLAC feature together with sub-space method for event (action) detection. For training and testing, both experiments use video data with multiple persons in the scene which are significantly different and rather complex as compared with the KTH datasets.

In the following, we compute a number of the CHLAC features in train and test intervals, N_h and M_h in the *h*-th hierarchy: $(N_0, M_0) = (10, 10)$, $(N_1, M_1) = (20, 20)$, $(N_2, M_2) = (30, 30)$, $(N_3, M_3) = (40, 40)$, $(N_4, M_4) = (50, 50)$, respectively in Figure 4.

Here, the hierarchical structure is determined heuristically by considering event category and time duration of that event. However, if we assume off-line processing, it is possible to search extensively hierarchical structure based on some coarse-to-fine strategy even if we have no information about event category and time duration of that event.

Using the hierarchical structure in Figure 4, the processing time necessary for the sequence of data with 300 frames, composed of 320 by 240[pixels] images, was about 300[msec] on PC with Intel Core2 Duo 2.53GHz and 2.0[GB] memory. Here, the processing time necessary for calculating CHLAC features in each frame can be almost ignored because of applying pipeline processing.

3.1. Detection of abnormal actions in *walking* scene

Typical examples of abnormal actions while walking are to crash with other persons and/or to tumble. As shown in Figure 5 and 6, here we have such scenes with multiple people walking in different directions. In this case, normal states are defined in the walking activities (Figure 5). Suppose we take sequences so that normal actions are included for the training datasets, $\{\mathbf{x}_j^{tr}\}_{j=1}^{M_h}$, and abnormal actions (Figure 6) in the test datasets, $\{\mathbf{x}_i^{te}\}_{i=1}^{N_h}$, both with time intervals, (M_h, N_h), as in Figure 4, then we find events as change of actions from normal state.

In the first experiment, we show the performance of proposed method for distinguishing abnormal actions (i.e. one person tumbles while other persons keep walking) from normal actions (i.e., every person keeps walking) in our in-house video sequences.

Figure 7 shows the time series data of anomaly (e.g., event likelihood) score *S* indicating a number of peaks with

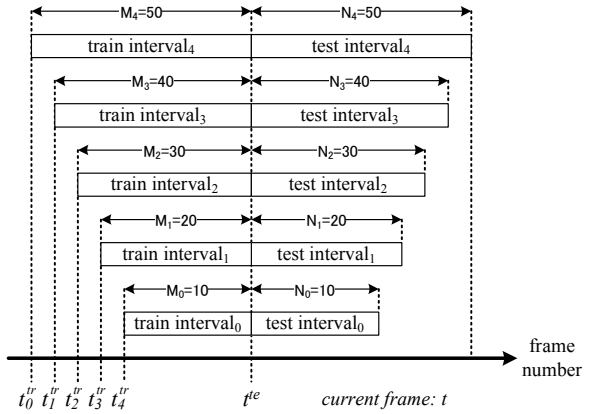

Figure 4: Hierarchical structure of time-scales for training and testing sequences.

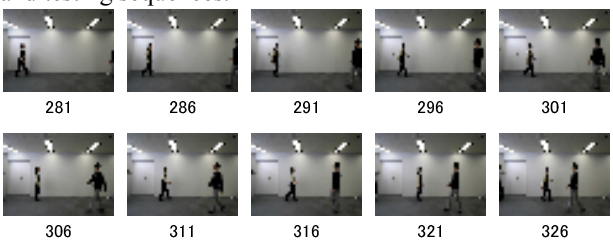

Figure 5: The scene of walking (normal state).

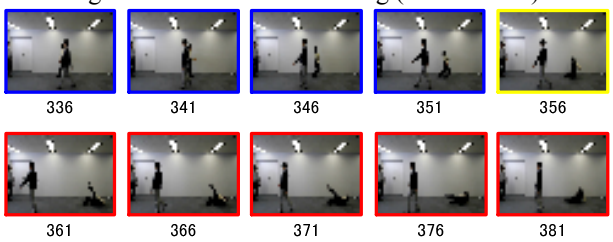

Figure 6: The scene for a person to tumble while other persons are walking.

scattered flat regions. From the 200-th frame to the 300-th frame in which the anomaly score *S* is relatively small corresponds to the walking scene (i.e. normal actions). On the other hand, from 300-th to 400-th frame, the anomaly score *S* is relatively high, corresponding to the tumbling scene (i.e., abnormal actions).

In Figure 6, the frame marked yellow (frame #356) indicates the frame with locally peaked score *S*, while frames marked blue (from frame #336 to #351) and red (from #361 to #381) are the preceding and succeeding sequences, respectively.

The result clearly shows that the proposed unsupervised method could detect the onset of abnormal actions and also distinguish normal from abnormal actions in this video scenario. For example, Figure 8 shows that abnormal actions (i.e., tumbling scene) were detected in the *5*-th hierarchy, relatively long interval: $(N_5, M_5) = (50, 50)$ which is appropriately interval of that event.

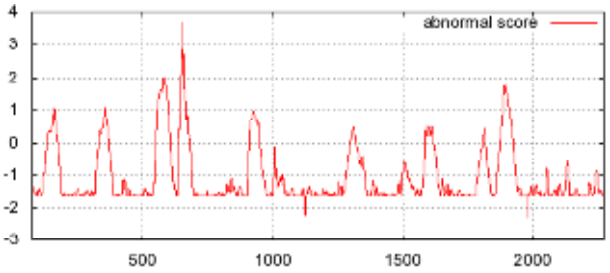

Figure 7: The sequential data of anomaly score *S*.

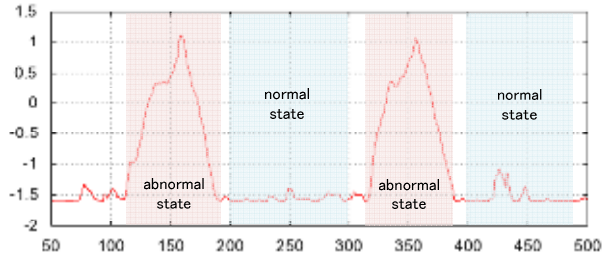

Figure 8: Enlargement of Figure 7; from the beginning to 500-th frame.

3.2. Detection of various actions in *tennis match* scenes

In this section, we show results obtained from tennis match video for action detection. We can observe various events such as ball person running in the court, routine actions (e.g., bouncing the ball) before service, and several smashing shots by player, which can be seen in Figure 11 - Figure 13. The video sequences have a wide spectrum of motion frequencies. In this case, normal states are defined in the walking actions (Figure 9).

Figure 14 shows sequential data of anomaly score *S* as applied to the tennis match scene. In Figure 15, from the beginning to 500-th frame, there are many peaks from the 150-th to the 200-th frame with relatively long intervals video sequences of ball person's running activities are given in Figure 10. Similarly, there is remarkable peak from the 250-th frame to the 350-th frame with relatively short intervals, and sequences that duration of 20 frames indicates routine action of server in Figure 11.

Figure 16 shows enlargement, from the 800-th frame to the 1300-th frame, with several peaks. Around the 900-th frame, for example, video sequences of 30 preceding/succeeding frames centered on the local maximum frame are the scene of backhand stroke in Figure 12. Similarly, around the 1000-th frame, (i.e., maximum interval is 20), are video sequences of 40 frames width around the peak frame corresponds to the scene of forehand stroke in Figure 13.

These results indicate successful detection of events as well as their onset frame. Especially, these results clearly show that events involving small actions with short time-scale such as routine motions are detected as ripples

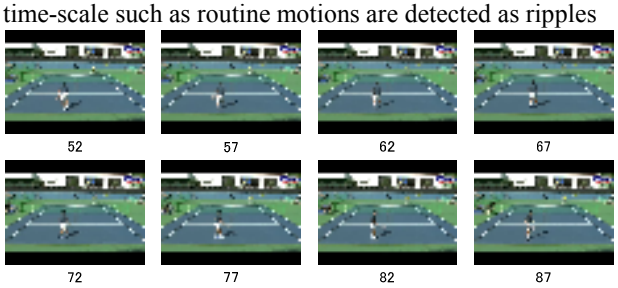

Figure 9: The scene of walking (normal state).

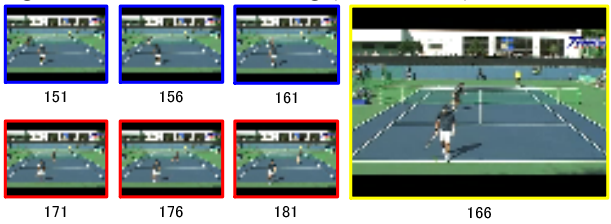

Figure 10: The scene of ball person's running.

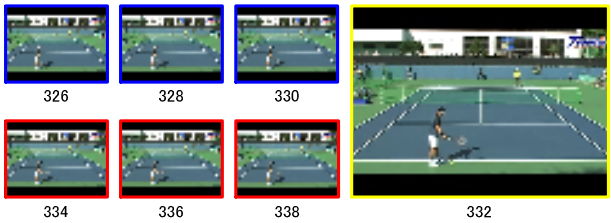

Figure 11: The scene of routine motions before service. Frame # 332 corresponds to just the middle of the scene.

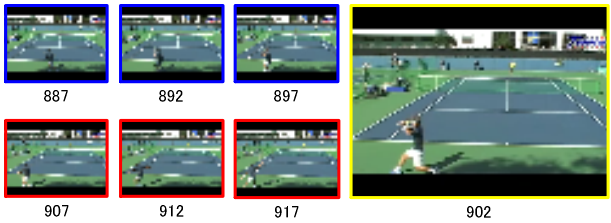

Figure 12: The scene of backhand stroke. The frame # 902 corresponds to the instance of smashing.

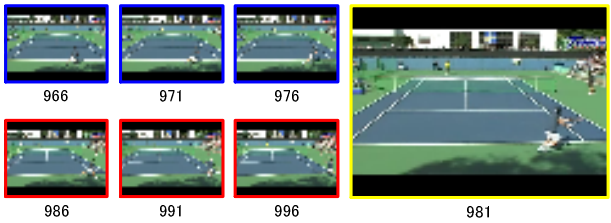

Figure 13: The scene of forehand stroke. The frame # 981 corresponds to the instance of smashing.

with shortest time scale (e.g., N_0 = M_0 =10 in Figure 4), whereas, events involving long duration actions such as ball person's actions are detected using a longer time interval (e.g., N_4=M_4=50 in Figure 4), in the hierarchical structure of time-scales for training and testing sequences (Figure 4).

3.3. Comparison with the reference method

Figure 17 shows sequential data of score S as applied to the same video sequences (i.e., walking scene) used in Section 3.1. The upper and lower parts are results obtained using proposed and reference method [10, 28], respectively. Scores of the reference method is given by sub-space method which represents the distance with normal space. The result clearly indicates that both methods give almost the same frame number for local maximum which corresponds to the event of person's tumbling.

However, the essential difference is that the proposed method finds both the mid frame of the event (i.e. the instance of tumbling) and its duration. Moreover, the proposed method can detect abnormal actions based on unsupervised approach by using the brief video data (on the order from 10 frames to 50 frames) as shown in Figure 5 and Figure 6. On the contrary, the reference method needs to learn the category of 'normal' action using sequence video data of 1000 frames beforehand, which are quite larger as compared with the proposed algorithm.

4. Summary and conclusions

We formulated the problem of video event detection as hierarchical, direct estimation of probability density ratios over two time intervals (e.g., past and current). Within this framework, we proposed a novel non-parametric algorithm for video event (i.e. actions/activities different from normal states) detection without categorization. Since the ratio of two probability densities is directly estimated without going through density estimation, the proposed method avoids nonparametric density estimation in event detection, which is known to be a hard problem in practice.

Based on the hierarchical framework with differing time-scales, we demonstrated detection of various video events with actions from micro level of small time-scale to macro level of larger time-scale. We experimentally showed the usefulness of the proposed method with two video sequences (i.e., walking scene and tennis match scene). First, we showed that the proposed unsupervised method could detect the onset of abnormal actions and also distinguish normal from abnormal actions in the two video scenarios. Second, we demonstrated that our method detects wide spectra of events as something significantly different from 'normal' states and also identifies their time intervals (e.g., N_h and M_h in the hierarchy of Figure 4) which cannot be obtained in the reference method [10]. Moreover, the proposed algorithm require extremely small datasets (e.g., on the order of 10 to 50 frames) in the training phase, whereas it is, in general, necessary to prepare huge amount of video data for learning action category (e.g., normal states) in conventional methods [1, 7, 10, 13, 15, 20]. For future work, we will extend the proposed framework to multi-category event detection with explicit categorization.

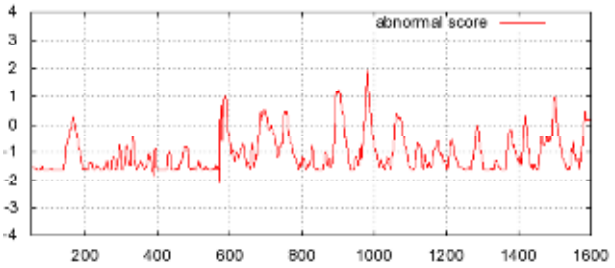

Figure 14: The sequential data of anomaly score *S* in the entire scene of tennis match.

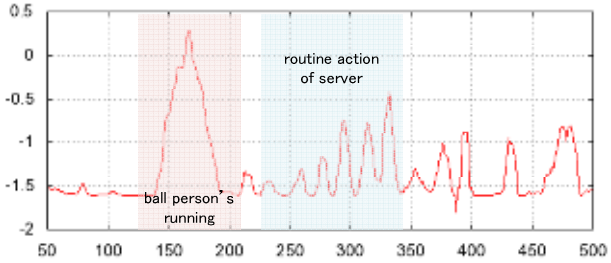

Figure 15: Enlargement of Figure 12; from the beginning to 500-th frame.

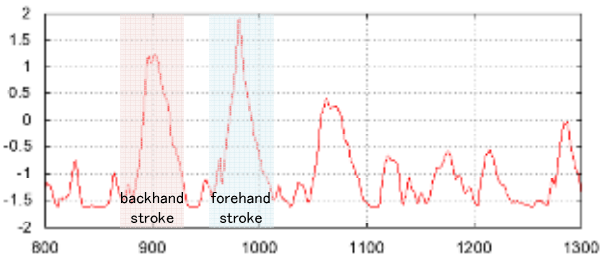

Figure 16: Enlargement of Figure 12; 800-1300-th frame.

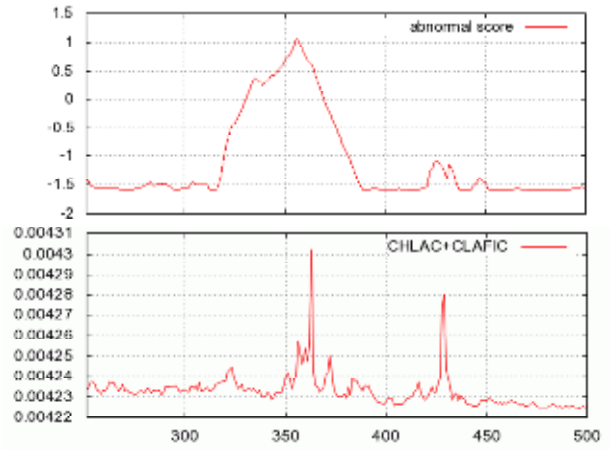

Figure 17: The sequential data of anomaly score *S* from the 251-th to 500-th frame; proposed method (upper part) and existing method (lower part).

References

[1] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, Event Detection and Recognition for Semantic Annotation of Video. *Multimedia Tools and Applications*, 51(1):279-302, 2011.

[2] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, Effective Codebooks for Human Action Categorization. In *Proc. of ICCV International Workshop on Video-oriented Object and Event Classification (VOEC)*, 2009.

[3] L. Ballan, M. Bertini, A. Del Bimbo, G. Serra, Video event classification using string kernels. *Multimedia Tools and Applications* 48(1):69-87, 2010.

[4] C. Chao, H. C. Shih, C. L. Huang Semantics-based highlight extraction of soccer program using DBN. In *Proc. of ICASSP,* 2005.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, Behavior recognition via sparse spatio-temporal features. In *Proc. of VSPETS*. 2005.

[6] S. Ebadollahi, L. Xie, S. F. Chang, J. Smith, Visual event detection using multi-dimensional concept dynamics. In *Proc. of Int'l Conference on Multimedia & Expo (ICME)* 2006.

[7] W. Jiang, C. Cotton, D. Ellis, and A. C. Loui, Short-Term Audio Visual Atoms for Generic Video Concept Classification, In *Proc. of ACM Multimedia*, 2009.

[8] L. Jie, B. Caputo, and V. Ferrari, Who's Doing What: Joint Modeling of Names and Verbs for Simultaneous Face and Pose Annotation, In *Proc. of NIPS*, 2009.

[9] T. Kanamori, S. Hido, and M. Sugiyama, Efficient Direct Density Ratio Estimation for Non-stationary Adaptation and Outlier Detection, In *Proc. of NIPS*, 2008.

[10] T. Kobayashi and N. Otsu, Action and Simultaneous Multiple-Person Identification Using Cubic Higher-Order Local Auto-Correlation, In *Proc. ICPR*, pp.741-744, 2004.

[11] I. Laptev, On space-time interest points. *International Journal of Computer Vision* 64(2-3):107-123, 2005.

[12] K. Mikolajczyk, and H. Uemura, Action Recognition with Motion-Appearance Vocabulary Forest. In *Proc. of CVPR*, 2008.

[13] J. Niebles, H. Wang, and L. Fei-Fei, Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words, *International Journal of Computer Vision*, 79(3):299-318, 2008.

[14] A. Oikonomopoulos, I. Patras, and M. Pantic, Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man and Cybernetics* 36(3):710-719, 2005.

[15] K. Rapantzikos, Y. Avrithis, and S. Kollias, Spatiotemporal Features for Action Recognition and Salient Event Detection, *Cognitive Computation* 3:167-184, 2011.

[16] H. J. Seo and P. Milanfar, Detection of Human Actions from a Single Example. In *Proc. of ICCV*, 1965-1970, 2009.

[17] G. Xu, Y. F. Ma, H.J. Zhang, and S. Yang, A HMM based semantic analysis framework for sports game event detection. In *Proc. of ICIP,* 2003.

[18] D. Xu, and S. F. Chang, Video Event Recognition Using Kernel Methods with Multilevel Temporal Alignment. *TPAMI* 30(11):1985-1997, 2008.

[19] B. Yao, and L. Fei-Fei, Grouplet: A Structured Image Representation for Recognizing Human and Object Interaction, In *Proc. of CVPR,* 2010.

[20] T.-H. Yu, and Y.-S. Moon, Unsupervised Abnormal Behavior Detection for Real-time Surveillance Using Observed History, In *Proc. of MVA2009 IAPR Conf. on Machine Vision Applications*, 166-169, 2009.

[21] X. Zhou, X. Zhuang, S. Yan, S. F. Chang, M. Hasegawa-Johnson, and T.S. Huang, SIFT-Bag Kernel for Video Event Analysis. In *Proc. of ACM Multimedia*, 2008.

[22] C. Wang, L. Zhang, and H.-J. Zhang, Learning to Reduce the Semantic Gap in Web Image Retrieval and Annotation. In *Proc. of ACM SIGIR*, 2008.

[23] R. Fujimaki, T. Yairi, and K. Machida, An Approach to Spacecraft Anomaly Detection Problem Using Kernel Feature Space, In *Proc. of ACM SIGKDD*, 2005.

[24] T. Kanamori, S. Hido, and M. Sugiyama, A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10, 1391-1445, 2009.

[25] T. Matsukawa, and T. Kurita, Action recognition using three-way cross-correlations feature of local motion attributes. In *Proc. of ICPR*, 1731-1734, 2010.

[26] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua. Contextual Bag of Words for Visual Categorization. *IEEE Trans. on Circuits and Systems for Video Technology*, 21(4):381-392, 2011.

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

[28] http://www.aist.go.jp/aist_e/latest_research/2005/20050621/20050621.html, World's Best Performance for Recognition of persons / individuals and movements / motions. *AIST press release,* 2005.