Regular Paper

Detection of Activities and Events without Explicit Categorization

Masao Yamanaka[1,2,a)] Masakazu Matsugu[2,b)] Masashi Sugiyama[3,c)]

Received: August 21, 2012, Revised: November 6, 2012,
Accepted: January 29, 2013

Abstract: We propose a method of unsupervised event detection from a video that compares probability distributions of past and current video sequence data in a sequential and hierarchical way. Because estimation of probability distributions is known to be difficult, naively comparing probability distributions via probability distribution estimation tends to be unreliable in practice. To cope with this problem, we use the state-of-the-art machine learning technique called *density ratio estimation*: The ratio of probability densities is directly estimated without density estimation, and thus probability distributions can be compared in a reliable way. Through experiments on a walking scene and a tennis match, we demonstrate the usefulness of the proposed approach.

Keywords: event detection, direct density-ratio estimation, cubic higher-order local auto-correlation

1. Introduction

Analysis of events and human actions from videos is useful in various applications such as content-based video retrieval, visual surveillance, and human-computer interaction. For this reason, event detection and human action recognition have gathered a great deal of attention recently [*1] [2], [17]. However, because of a wide variety of backgrounds, contexts, and events, these tasks are known to be highly challenging [12]. Furthermore, considerable variations in clothes, sizes, or postures of people, illumination conditions, occlusion conditions, and camera angels could make the tasks even more difficult.

Most event detection methods first extract primitive features from video sequences such as optical flow based features [15], [27], spatio-temporal features [4], [9], [10], [11], [15], [17], [18], [22], or static features including appearance, shape, and spatial relations among local features [7], [23]. Some approaches further utilize more sophisticated codebook representation [1], [16], [26] that is effective for describing and discriminating between various event categories. In particular, the bag-of-words representation of spatio-temporal points-of-interest has received considerable attention [1], [12], [16], [28]. Then these features are fed into learning machines such as hidden Markov models [5], [24], Bayesian networks [3], kernel methods [4], [23], [28], and tree-structured classifiers [15] to recognize events and actions.

A standard approach involves supervised training of a classifier on a huge amount of video data with ground truth annotations [6]. However, because gathering annotated data is costly, it is often difficult to gather an enough amount of annotated data in practice. To cope with this problem, unsupervised approaches have also been actively explored recently for learning human action categories [15], [19] and for detecting abnormal behavior [27]. However, achieving higher accuracy by the unsupervised approach is still highly challenging.

In this paper, we propose a new unsupervised detection method of semantic event categories that compares probability distributions of past and current video data. However, estimation of probability distributions is known to be difficult [21], and thus distribution comparison via probability distribution estimation tends to be unreliable in practice. To cope with this problem, we use the state-of-the-art machine learning technique called *density ratio estimation* [20], which avoids distribution estimation and directly estimates the ratio of probability densities. Thus, distribution comparison via density ratio estimation is more reliable than the naive approach based on probability distribution estimation [8]. We sequentially compare the probability distributions of past and current sequence data in various time scales described by spatio-temporal features (more specifically, we use the *cubic higher-order local auto-correlation*; CHLAC [10]). The usefulness of the proposed method is demonstrated through experiments on a walking scene and a tennis match.

The remainder of this paper is structured as follows. In Section 2, we describe our event detection method based on density ratio estimation. In Section 3, we report experimental results on real-world video sequences. Finally, we conclude by summarizing our contributions in Section 4.

[1] Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Kanagawa 226–8502, Japan
[2] CANON Inc., Ohta, Tokyo 146–8501, Japan
[3] Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Meguro, Tokyo 152–8552, Japan
[a)] yamanaka@sp.dis.titech.ac.jp
[b)] matsugu.masakazu@canon.co.jp
[c)] sugi@cs.titech.ac.jp

[*1] In this paper, we regard "events" as changes of human activities that are observed as visual contents.

2. Problem Formulation and Proposed Approach

In this section, we formulate the event detection problem based on density ratio estimation, and describe our proposed approach.

2.1 Problem Formulation

Let $\boldsymbol{x}(t)$ be a spatio-temporal feature vector at time t (we use 251-dimensional CHLAC features [10] in our experiments; see Section 3.1 for details). Our task is to detect whether there exists a change point between two consecutive time intervals called the *train* and *test* intervals (see **Fig. 1**, where M and N denote the numbers of frames in the train and test intervals, respectively).

Let $p_{\mathrm{tr}}(\boldsymbol{x})$ and $p_{\mathrm{te}}(\boldsymbol{x})$ be the probability density functions of the train and test time-series features, respectively. A naive approach to evaluating the difference between train and test intervals would be to first estimate the train and test probability density functions separately from the train and test time-series data, and then compare the estimated probability densities. However, since non-parametric density estimation is known to be a hard problem [21], this naive approach may not be effective in practice. Instead, directly estimating the *ratio* of probability densities without going through density estimation was shown to be more promising [20].

Motivated by this line of research, we develop an event detection algorithm based on the *relative density ratio* [25] for spatio-temporal feature $\boldsymbol{x}$:

$$w(\boldsymbol{x}) = \frac{p_{\mathrm{tr}}(\boldsymbol{x})}{\beta p_{\mathrm{tr}}(\boldsymbol{x}) + (1-\beta)p_{\mathrm{te}}(\boldsymbol{x})}, \qquad (1)$$

where $p_{\mathrm{tr}}(\boldsymbol{x})$ and $p_{\mathrm{te}}(\boldsymbol{x})$ are the probability density functions for train and test spatio-temporal features, respectively. β $(0 \leq \beta < 1)$ is the relative parameter that controls the "smoothness" of the density ratio; $\beta = 0$ corresponds to the plain density ratio $p_{\mathrm{tr}}(\boldsymbol{x})/p_{\mathrm{te}}(\boldsymbol{x})$ and the relative density ratio tends to be smoother as β gets larger. See [25] for a theoretical background for this relative parameter. Based on the relative density ratio, we define our anomaly score as the *relative Pearson divergence* from $p_{\mathrm{tr}}(\boldsymbol{x})$ to $p_{\mathrm{te}}(\boldsymbol{x})$ [13], [25]:

$$\frac{1}{2}\int \Big(w(\boldsymbol{x}) - 1\Big)^2 \Big(\beta p_{\mathrm{tr}}(\boldsymbol{x}) + (1-\beta)p_{\mathrm{te}}(\boldsymbol{x})\Big)\,\mathrm{d}\boldsymbol{x},$$

which is always non-negative and zero if and only if $p_{\mathrm{tr}} = p_{\mathrm{te}}$.

In practice, it may be difficult to determine the size of train and test intervals to properly detect events without any prior knowledge. In this paper, we mitigate this difficulty by considering a hierarchy of train and test intervals, as illustrated in **Fig. 2**. This hierarchical structure makes it possible to detect events in different time scales from micro to macro levels. Let S_h be the anomaly score in the h-th hierarchy. Then the final anomaly score S is defined by

$$S = \max_h S_h.$$

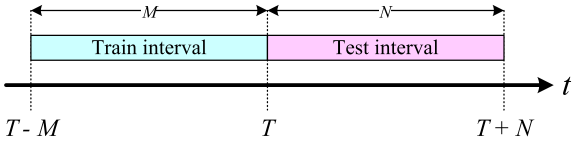

Fig. 1 Definition of train and test intervals.

The remaining question in the proposed procedure is how to accurately evaluate the relative Pearson divergence from data, which is explained below.

2.2 Approximation of Relative Pearson Divergence by Relative Density-ratio Estimation

Suppose that we are given a set of N_{tr} samples extracted from a train interval drawn independently from a probability distribution P_{tr} with density p_{tr}:

$$\mathcal{X}_{\mathrm{tr}} = \left\{\boldsymbol{x}_i^{\mathrm{tr}} \mid \boldsymbol{x}_i^{\mathrm{tr}} \in \Re^d\right\}_{i=1}^{N_{\mathrm{tr}}} \overset{\mathrm{i.i.d}}{\sim} P_{\mathrm{tr}}.$$

We also suppose that another set of N_{te} samples extracted from a test interval drawn independently from another probability distribution P_{te} with density p_{te} is available:

$$\mathcal{X}_{\mathrm{te}} = \left\{\boldsymbol{x}_j^{\mathrm{te}} \mid \boldsymbol{x}_j^{\mathrm{te}} \in \Re^d\right\}_{j=1}^{N_{\mathrm{te}}} \overset{\mathrm{i.i.d}}{\sim} P_{\mathrm{te}}.$$

From the samples $\mathcal{X}_{\mathrm{tr}}$ and $\mathcal{X}_{\mathrm{te}}$, we approximate the Pearson divergence. If an estimator of the relative density ratio, $\widehat{w}(\boldsymbol{x})$, is available, the Pearson divergence can be approximated as

$$-\frac{\beta}{2N_{\mathrm{tr}}}\sum_{i=1}^{N_{\mathrm{tr}}} \widehat{w}(\boldsymbol{x}_i^{\mathrm{tr}})^2 - \frac{1-\beta}{2N_{\mathrm{te}}}\sum_{j=1}^{N_{\mathrm{te}}} \widehat{w}(\boldsymbol{x}_j^{\mathrm{te}})^2 + \frac{1}{N_{\mathrm{tr}}}\sum_{i=1}^{N_{\mathrm{tr}}} \widehat{w}(\boldsymbol{x}_i^{\mathrm{tr}}) - \frac{1}{2}.$$

Below, we explain an estimation method of relative density ratios called *relative unconstrained least-squares importance fitting* (RuLSIF) [25].

Let us model the relative density ratio function $w(\boldsymbol{x})$ by the following kernel model:

$$\widetilde{w}(\boldsymbol{x}) = \sum_{i=1}^{N_{\mathrm{tr}}} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i^{\mathrm{tr}}) = \boldsymbol{\alpha}'\boldsymbol{k}(\boldsymbol{x}),$$

where

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{N_{\mathrm{tr}}})'$$

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

$$\boldsymbol{k}(\boldsymbol{x}) = \Big(K(\boldsymbol{x}, \boldsymbol{x}_1^{\mathrm{tr}}), K(\boldsymbol{x}, \boldsymbol{x}_2^{\mathrm{tr}}), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_{N_{\mathrm{tr}}}^{\mathrm{tr}})\Big)'$$

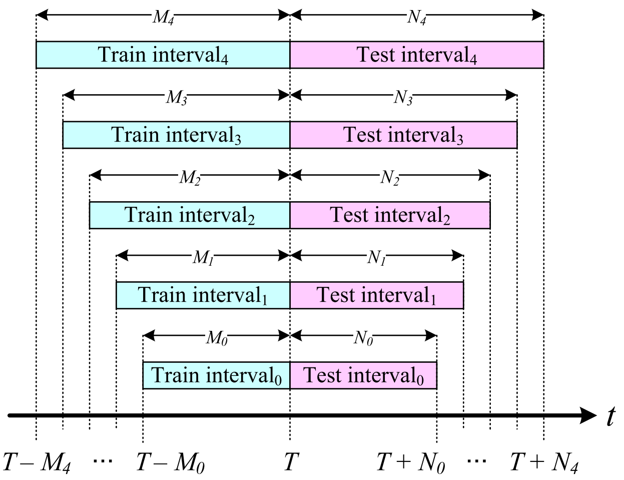

Fig. 2 Definition of train and test intervals in a hierarchy.

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

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

where $\sigma > 0$ is the Gaussian width.

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

$$\begin{aligned} J_0 &= \frac{1}{2}\int \left(\widetilde{w}(\boldsymbol{x}) - w(\boldsymbol{x})\right)^2 \left(\beta p_{\mathrm{tr}}(\boldsymbol{x}) + (1-\beta)p_{\mathrm{te}}(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x} \\ &= \frac{\beta}{2}\int \widetilde{w}(\boldsymbol{x})^2 p_{\mathrm{tr}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \frac{1-\beta}{2}\int \widetilde{w}(\boldsymbol{x})^2 p_{\mathrm{te}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &\quad - \int \widetilde{w}(\boldsymbol{x}) p_{\mathrm{tr}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \mathrm{Const.} \end{aligned}$$

Let us denote the first three terms by J. Since J contains the expectation over unknown densities $p_{\mathrm{te}}(\boldsymbol{x})$ and $p_{\mathrm{tr}}(\boldsymbol{x})$, we approximate the expectations by empirical averages. Then we obtain

$$\widehat{J}(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}'\widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \boldsymbol{\alpha}'\widehat{\boldsymbol{h}},$$

where $\widehat{\boldsymbol{H}}$ is the $N_{\mathrm{tr}} \times N_{\mathrm{tr}}$ matrix defined by

$$\widehat{\boldsymbol{H}} = \frac{\beta}{N_{\mathrm{tr}}}\sum_{i=1}^{N_{\mathrm{tr}}} \boldsymbol{k}(\boldsymbol{x}_i^{\mathrm{tr}})\boldsymbol{k}(\boldsymbol{x}_i^{\mathrm{tr}})' + \frac{1-\beta}{N_{\mathrm{te}}}\sum_{j=1}^{N_{\mathrm{te}}} \boldsymbol{k}(\boldsymbol{x}_j^{\mathrm{te}})\boldsymbol{k}(\boldsymbol{x}_j^{\mathrm{te}})',$$

and $\widehat{\boldsymbol{h}}$ is the N_{tr}-dimensional vector defined by

$$\widehat{\boldsymbol{h}} = \frac{1}{N_{\mathrm{tr}}}\sum_{i=1}^{N_{\mathrm{tr}}} \boldsymbol{k}(\boldsymbol{x}_i^{\mathrm{tr}}).$$

By including a regularization term, the RuLSIF optimization problem is formulated as

$$\widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\mathrm{argmin}}\left[\frac{1}{2}\boldsymbol{\alpha}'\widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \boldsymbol{\alpha}'\widehat{\boldsymbol{h}} + \frac{\lambda}{2}\boldsymbol{\alpha}'\boldsymbol{\alpha}\right],$$

where $\boldsymbol{\alpha}'\boldsymbol{\alpha}/2$ is a regularizer and λ (≥ 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 $\widehat{\boldsymbol{\alpha}}$ as

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I})^{-1}\widehat{\boldsymbol{h}},$$

where $\boldsymbol{I}$ denotes the identity matrix. Finally, a density ratio estimator $\widehat{w}(\boldsymbol{x})$ is given by

$$\widehat{w}(\boldsymbol{x}) = \widehat{\boldsymbol{\alpha}}'\boldsymbol{k}(\boldsymbol{x}).$$

Thanks to the simple analytic-form expression, RuLSIF is computationally efficient.

2.3 Model Selection by Cross-validation

The practical performance of RuLSIF 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 RuLSIF is possible based on *cross-validation* with respect to the error criterion J.

More specifically, each of the sample sets $\mathcal{X}_{\mathrm{tr}} = \{\boldsymbol{x}_i^{\mathrm{tr}}\}_{i=1}^{N_{\mathrm{tr}}}$ and $\mathcal{X}_{\mathrm{te}} = \{\boldsymbol{x}_j^{\mathrm{te}}\}_{j=1}^{N_{\mathrm{te}}}$ is divided into L disjoint sets $\{\mathcal{X}_{\mathrm{tr}}^l\}_{l=1}^L$ and $\{\mathcal{X}_{\mathrm{tr}}^l\}_{l=1}^L$. Then a RuLSIF solution $\widetilde{w}_l(\boldsymbol{x})$ is obtained using $\mathcal{X}_{\mathrm{tr}}\backslash\mathcal{X}_{\mathrm{tr}}^l$ and $\mathcal{X}_{\mathrm{te}}\backslash\mathcal{X}_{\mathrm{te}}^l$ (i.e., all samples without $\mathcal{X}_{\mathrm{tr}}^l$ and $\mathcal{X}_{\mathrm{te}}^l$), and its J-value for the hold-out samples $\mathcal{X}_{\mathrm{tr}}^l$ and $\mathcal{X}_{\mathrm{te}}^l$ is computed as

$$\begin{aligned} \widehat{J}_{\mathrm{CV}}^l &= \frac{\beta}{2|\mathcal{X}_{\mathrm{tr}}^l|}\sum_{\boldsymbol{x}_{\mathrm{tr}}\in\mathcal{X}_{\mathrm{tr}}^l} \widetilde{w}_l(\boldsymbol{x}_{\mathrm{tr}})^2 + \frac{1-\beta}{2|\mathcal{X}_{\mathrm{te}}^l|}\sum_{\boldsymbol{x}_{\mathrm{te}}\in\mathcal{X}_{\mathrm{te}}^l} \widetilde{w}_l(\boldsymbol{x}_{\mathrm{te}})^2 \\ &\quad - \frac{1}{|\mathcal{X}_{\mathrm{tr}}^l|}\sum_{\boldsymbol{x}_{\mathrm{tr}}\in\mathcal{X}_{\mathrm{tr}}^l} \widetilde{w}_l(\boldsymbol{x}_{\mathrm{tr}}), \end{aligned}$$

where $|\mathcal{X}|$ denotes the number of elements in the set $\mathcal{X}$. This procedure is repeated for $l = 1, \ldots, L$, and the average of $\widehat{J}_{\mathrm{CV}}^l$ over all l is computed as

$$\widehat{J}_{\mathrm{CV}} = \frac{1}{L}\sum_{l=1}^{L}\widehat{J}_{\mathrm{CV}}^l.$$

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

3. Experiments

In this section, we show two experimental studies on a walking scene and a tennis match to evaluate the performance of the proposed method. Through all experiments, we set the relative parameter at $\beta = 0.1$.

3.1 Cubic Higher-order Local Auto-correlation

In our experiments, we adopt the spatio-temporal features called the *cubic higher-order local auto-correlation* (CHLAC) [10], which and whose extension [14] have been successfully used in action recognition. CHLAC directly handles three-dimensional data, and it possesses useful properties such as additivity, shift invariance, and robustness to noise [10].

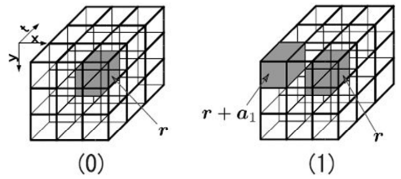

Fig. 3 Example of a mask pattern. (0) $L = 0$. (1) $L = 1$ for $a_1 = (-1, -1, -1)$. Mutually shifted mask patterns are eliminated.

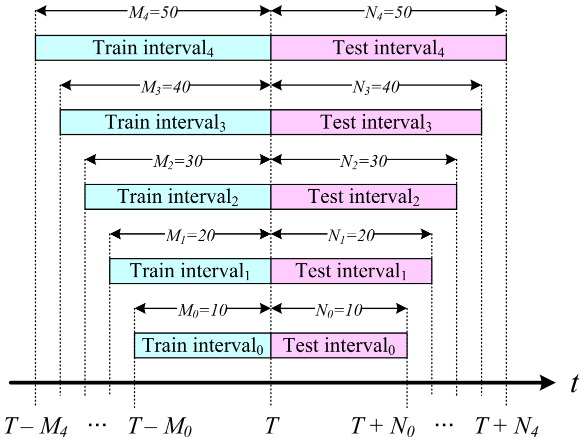

Fig. 4 Hierarchical structure of time scales for train and test sequences.

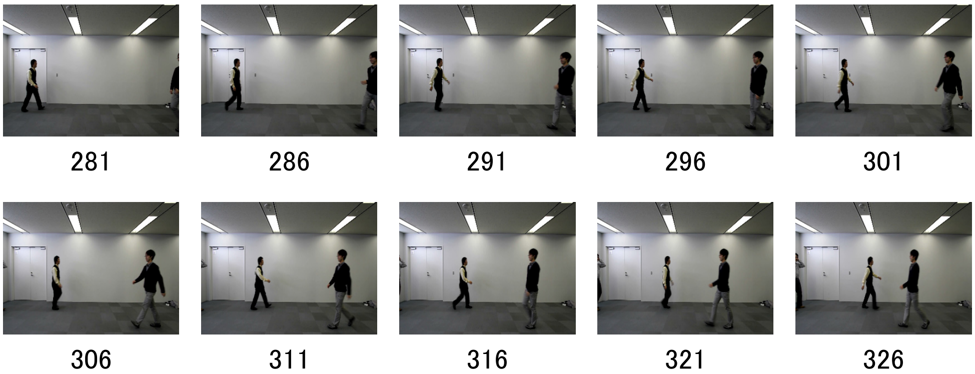

Fig. 5 An ordinary walking scene.

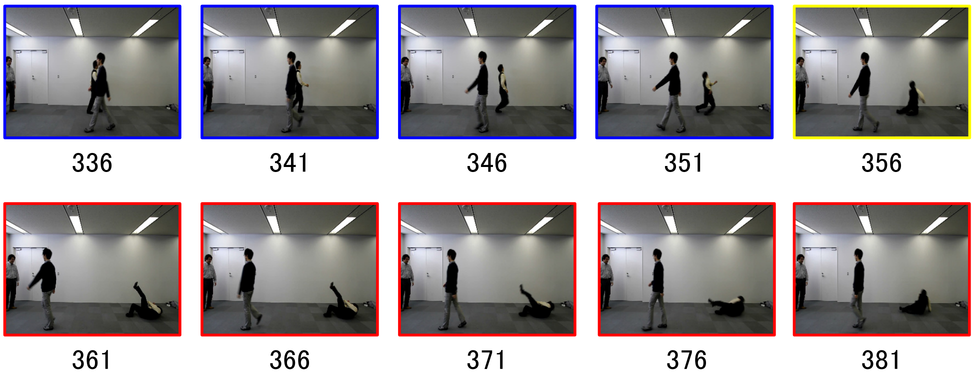

Fig. 6 A person tumbles while other people are walking.

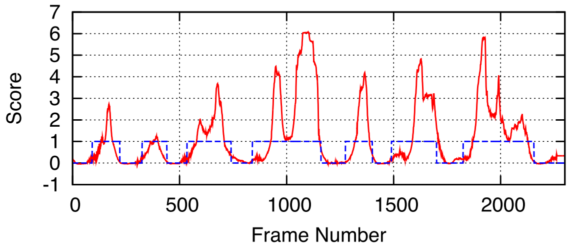

Fig. 7 Anomaly score obtained by the proposed method and the ground truths for the walking scene.

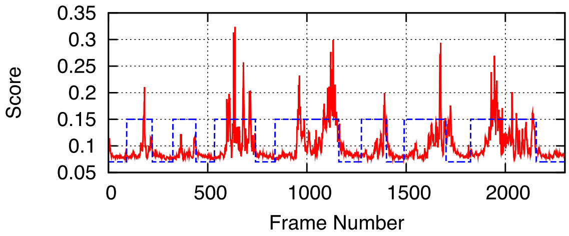

Fig. 9 Anomaly score obtained by the subspace method and the ground truths for the walking scene.

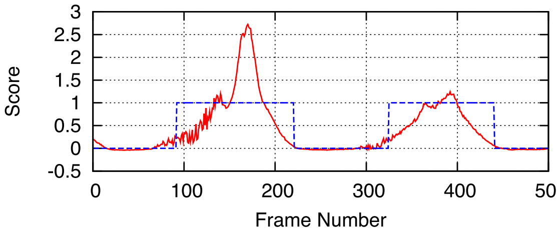

Fig. 8 Enlargement of Fig. 7 from the beginning to 500th frames.

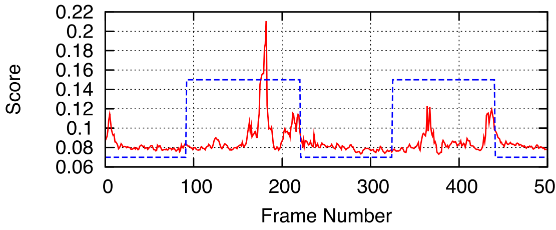

Fig. 10 Enlargement of Fig. 9 from the beginning to 500th frames.

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

$$\boldsymbol{x}_L(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_L) = \int f(\boldsymbol{r}) f(\boldsymbol{r} + \boldsymbol{a}_1) \cdots f(\boldsymbol{r} + \boldsymbol{a}_L) \mathrm{d}\boldsymbol{r}, \qquad (3)$$

where $\boldsymbol{a}_l$ $(l = 1, \ldots, L)$ are called the displacement vectors. In Eq. (3), displacement vector $\boldsymbol{a}_l$ is limited to a $3 \times 3 \times 3$ local region around $\boldsymbol{r}$ and the number of displacement vectors, L, is set to be less than or equal to 2.

Because the value of $\boldsymbol{x}_L(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_L)$ remains the same as long as the patterns of $(\boldsymbol{r}, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_L)$ are identical in the point configuration, we eliminate such redundant features (see **Fig. 3**). Taking inter-frame difference and thresholding, we convert input im-

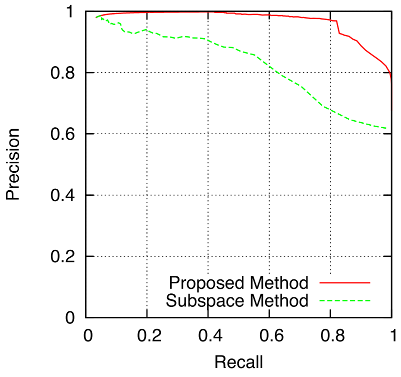

Fig. 11 Precision-recall curves.

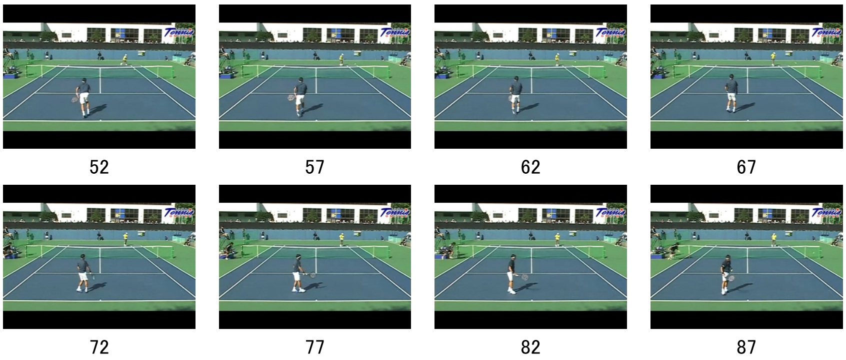

Fig. 12 A tennis match video.

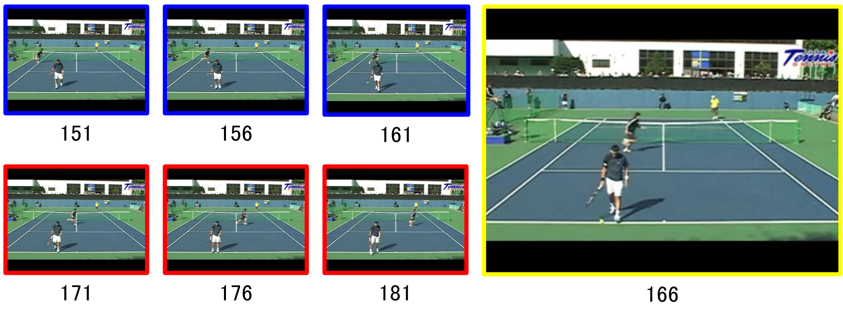

Fig. 13 A ball person is running in the court.

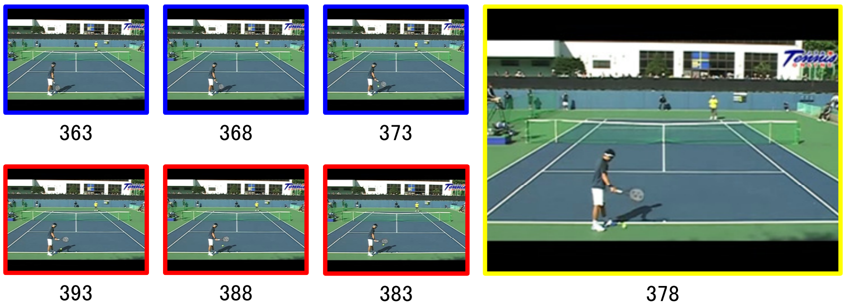

Fig. 14 A player bounces a ball before his service.

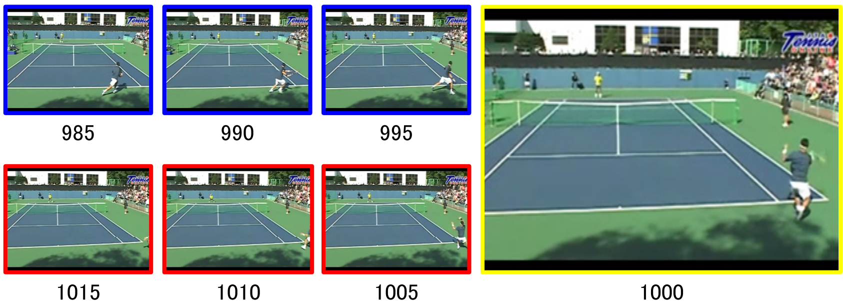

Fig. 15 A player makes a forehand smashing shot.

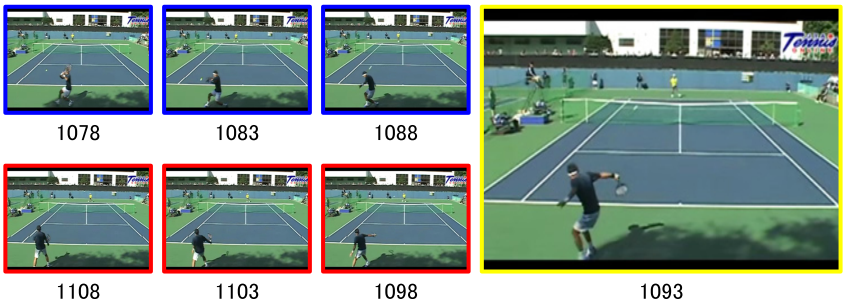

Fig. 16 A player makes a backhand smashing shot.

age data into motion-image sequences composed of binary data. We use 251 mask patterns, so our CHLAC feature vector is 251-dimensional.

As shown in **Fig. 4**, we compute CHLAC features in train and test intervals for $(M_0, N_0) = (10, 10)$, $(M_1, N_1) = (20, 20)$, $(M_2, N_2) = (30, 30)$, $(M_3, N_3) = (40, 40)$, and $(M_4, N_4) = (50, 50)$.

3.2 Detection of Abnormal Actions in Walking Scene

In the first set of experiments, we use our in-house video sequences that record walking scenes. While people are walking (**Fig. 5**), typical abnormal actions are that people are hit each other and/or tumble (**Fig. 6**).

We compare the performance of the proposed method with that of a baseline approach based on the subspace method [10]. The subspace method is a supervised approach and we pre-trained the subspace classifier using video data of 500 frames that have been manually annotated. On the other hand, the proposed method is completely unsupervised and no pre-training is necessary.

Figures 7 and **8** show the anomaly scores obtained by the proposed method, while **Figs. 9** and **10** show the anomaly scores obtained by the subspace method. In the graphs, the blue lines represent the manually annotated ground truths where normal periods (i.e., walking scene) are indicated by smaller values and abnormal periods (i.e., tumbling scene) are indicated by larger values. The scene from the 200th to 300th frames has relatively small anomaly scores and this corresponds to an ordinary walking scene. On the other hand, the scene from 300th to 400th frames has relatively high anomaly scores and this corresponds to a tumbling scene (see Fig. 6 for details). Both the proposed and existing methods successfully detect the onset of this abnormal action and also distinguish normal actions from abnormal ones.

Figure 11 plots precision-recall curves of abnormal action recognition. This clearly show that our proposed method outperforms the subspace method. The superior performance of the proposed method would be brought by the fact that direct density-ratio estimation is capable of comparing probability distributions in a highly robust manner.

Overall, the proposed method, which is an unsupervised method that does not require any pre-training was shown to compare favorably with the supervised subspace method that requires manually annotated video data.

3.3 Detection of Various Actions in Tennis Match

Next, we show experimental results on a tennis match video (see **Fig. 12**). In this video, various events can be observed, for example, a ball person is running in the court (**Fig. 13**), a player bounces a ball before his service (**Fig. 14**), and a player makes smashing shots (**Figs. 15** and **16**).

Figures 17, **18**, **19** show the anomaly score obtained by the proposed method. The peak between the 150th and 200th frames corresponds to a ball person's running in the court (Fig. 13), while the peak between the 300th and 400th frames corresponds to ball bouncing (Fig. 14). The peak around the 1000th frame corresponds to the forehand smashing stroke (Fig. 15), while the peak around the 1100th frame corresponds to the backhand smashing stroke (Fig. 16).

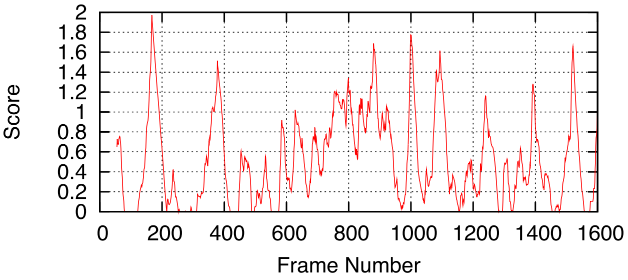

Fig. 17 Anomaly score obtained by the proposed method for the tennis match video.

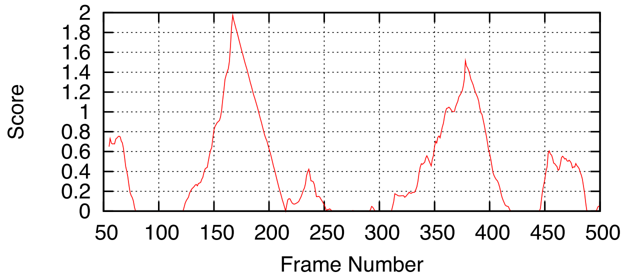

Fig. 18 Enlargement of Fig. 17 from the beginning to 500th frames.

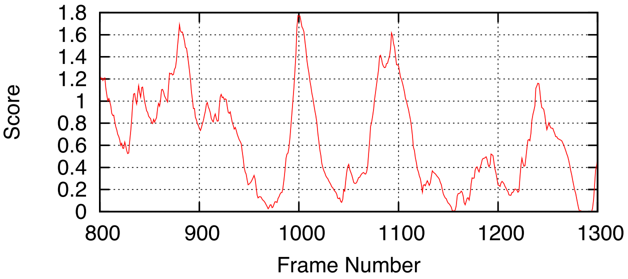

Fig. 19 Enlargement of Fig. 17 between the 800th and 1300th frames.

These results indicate that the proposed method can successful detect notable events as well as their onset frames.

4. Summary and Conclusions

We proposed a video-based event detection method using direct density-ratio estimation. A similar idea has already been explored in terms of change detection in time-series [8], [13], but we newly introduced a multi-scale hierarchy of train and test intervals—this mitigates the difficulty of finding appropriate time intervals. We experimentally demonstrated the usefulness of the proposed method on a walking scene and a tennis match. Because manually annotating video sequence is highly costly. the fact that the proposed algorithm does not require any pre-training based on annotated data is a significant advantage over existing supervised approaches in practice.

References

[1] Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L. and Serra, G.: Effective codebooks for human action categorization, *Proc. IEEE 12th International Conference on Computer Vision and Pattern Recognition*, pp.506–513 (2009).

[2] Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L. and Serra, G.: Event detection and recognition for semantic annotation of video, *Multimedia Tools and Applications*, Vol.51, No.1, pp.279–302 (2011).

[3] Ballan, L., Bertini, M., Del Bimbo, A. and Serra, G.: Video event classification using string kernels, *Multimedia Tools and Applications*, Vol.48, No.1, pp.69–87 (2010).

[4] Chao, C., Shih, H.C. and Huang, C.L.: Semantics-based highlight extraction of soccer program using dbn, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.1057–1060 (2005).

[5] Ebadollahi, S., Xie, L., Chang, S.F. and Smith, J.: Visual event detection using multi-dimensional concept dynamics, *Proc. IEEE International Conference on Multimedia and Expo*, pp.881–884 (2006).

[6] Jiang, W., Cotton, C., Ellis, D. and Loui, A.C.: Short-term audio-visual atoms for generic video concept classification, *Proc. 17th ACM International Conference on Multimedia*, pp.5–14 (2009).

[7] Jie, L., Caputo, B. and Ferrari, V.: Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation, *Advances in Neural Information Processing Systems 22*, pp.1168–1176 (2009).

[8] Kawahara, Y. and Sugiyama, M.: Sequential change-point detection based on direct density-ratio estimation, *Statistical Analysis and Data Mining*, Vol.5, No.2, pp.114–127 (2012).

[9] Ke, Y., Sukthankar, R. and Hebert, M.: Event detection in crowded videos, *Proc. IEEE 11th International Conference on Computer Vision*, pp.1–8 (2007).

[10] Kobayashi, T. and Otsu, N.: Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation, *Proc. 17th International Conference on Pattern Recognition*, pp.741–744 (2004).

[11] Laptev, I.: On space-time interest points, *International Journal of Computer Vision*, Vol.64, No.2, pp.107–123 (2005).

[12] Li, T., Mei, T., Kweon, I.-S. and Hua, X.-S.: Contextual bag-of-words for visual categorization, *IEEE Trans. Circuits and Systems for Video Technology*, Vol.21, No.4, pp.381–392 (2011).

[13] Liu, S., Yamada, M., Collier, N. and Sugiyama, M.: Change-point detection in time-series data by relative density-ratio estimation, Technical Report 1203.0453, arXiv (2012).

[14] Matsukawa, T. and Kurita, T.: Action recognition using three-way cross-correlations feature of local motion attributes, *Proc. 20th International Conference on Pattern Recognition*, pp.1731–1734 (2010).

[15] Mikolajczyk, K. and Uemura, H.: Action recognition with motion-appearance vocabulary forest, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.1–8 (2008).

[16] Niebles, J., Wang, H. and Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words, *International Journal of Computer Vision*, Vol.79, No.3, pp.299–318 (2008).

[17] Oikonomopoulos, A., Patras, I. and Pantic, M.: Spatiotemporal salient points for visual recognition of human actions, *IEEE Trans. Systems, Man, and Cybernetics*, Vol.36, No.3, pp.710–719 (2005).

[18] Rapantzikos, K., Avrithis, Y. and Kollias, S.: Spatiotemporal features for action recognition and salient event detection, *Cognitive Computation*, Vol.3, No.1, pp.167–184 (2011).

[19] Seo, H.J. and Milanfar, P.: Detection of human actions from a single example, *Proc. IEEE 12th International Conference on Computer Vision*, pp.1965–1970 (2009).

[20] Sugiyama, M., Suzuki, T. and Kanamori, T.: *Density Ratio Estimation in Machine Learning*, Cambridge University Press, Cambridge, UK (2012).

[21] Vapnik, V.N.: *Statistical Learning Theory*, Wiley, New York, NY, USA (1998).

[22] Wang, C., Zhang, L. and Zhang, H.J.: Learning to reduce the semantic gap in web image retrieval and annotation, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.355–362 (2008).

[23] Xu, D. and Chang, S.F.: Video event recognition using kernel methods with multilevel temporal alignment, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.30, No.11, pp.1985–1997 (2008).

[24] Xu, G., Ma, Y.F., Zhang, H.J. and Yang, S.: A hmm based semantic analysis framework for sports game event detection, *Proc. International Conference on Image Processing*, pp.25–28 (2003).

[25] Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H. and Sugiyama, M.: Relative density-ratio estimation for robust distribution comparison, *Advances in Neural Information Processing Systems 24*, pp.594–602 (2011).

[26] Yao, B. and Fei-Fei, L.: Grouplet: A structured image representation for recognizing human and object interaction, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.9–16 (2010).

[27] Yu, T.-H. and Moon, Y.-S.: Unsupervised abnormal behavior detection for real-time surveillance using observed history, *Proc. IAPR Conference on Machine Vision Applications*, pp.9–16 (2010).

[28] Zhou, X., Zhuang, X., Yan, S., Chang, S.F., Hasegawa-Johnson, M. and Huang, T.S.: SIFT-Bag kernel for video event analysis, *Proc. 16th ACM International Conference on Multimedia*, pp.229–238 (2008).

Masao Yamanaka has worked for corporate R&D Headquaters, Canon Inc. He has been engaged in the development of image information processing technology, such as face recognition in still image, pose estimation in depth image, and anomaly detection of human actions in video so far. His recent research interests include not only a wide range of image information processing techniques but also bioinformatics using advanced machine learning.

Masakazu Matsugu has worked on pattern recognition and neural networks. He received 2002 ICONIP Best Paper Award and 2003 FIT Outstanding Paper Award. He realized a visual inspection system in industrial production system and face identification as well as facial expression functionalities in digital cameras. He holds 110 Japanese patents and 95 US patents. He is a member of INNS.

Masashi Sugiyama received his B.E., M.E., and Ph.D. degrees from Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, in 1997, 1999, and 2001. In 2001, he was appointed as a Research Associate in the same institute, and from 2003, he is an Associate Professor. His research interests include theory and application of machine learning and signal/image processing.