Frequent Pattern Mining in Multiple Trajectories of Football Players

Yuto Suzuki and Tomonobu Ozaki

***Abstract*—Recently, it has been regarded as important in many sports fields to evaluate the tactics and athletes using actual play records. In this paper, as a first step towards a quantitative evaluation of the strategies in football games, we propose an algorithm for discovering frequent patterns on simultaneous trajectories of multiple football players. In the algorithm, given trajectories are firstly converted into a set of labeled sub-trajectories corresponding to the interval-based events. A pattern enumeration algorithm is then applied to the obtained interval-based events with a consideration of the order of events, the time difference and the spatial spread of sub-trajectories. We introduce variables for subjects of events (sub-trajectories) in the pattern. By using variables, we can recognize which events were played by the same player and which events were played differently. In addition, it is possible to extract a pattern which absorb the difference of concrete players. To evaluate the proposed algorithm, we conduct experiments using real trajectory datasets on nine matches in Japanese professional football league. The results on the computation time and the number of extracted patterns show the feasibility and effectiveness of the algorithm. In addition, we succeeded in extracting meaningful patterns representing certain offensive and defensive strategies formed by multiple football players.**

***Index Terms*—trajectory mining, frequent patterns, sequential patterns, football**

I. Introduction

In recent years, in various sports fields typified by volleyball, basketball and football, it is widely recognized as important to analyze the ability of the players and the team tactics based on not experience or intuition, but on the actually recorded play data. To evaluate the abilities quantitatively, various studies have been conducted. In the field of football for example, Kang[1] reports a quantitative evaluation of athletes' performance and Lucey[2] analyses the difference in tactics between home matches and away matches. In addition, Bialkowski[3] proposes an unsupervised method to learn a formation template.

In this paper, as a first step towards a quantitative evaluation of the tactical movements among multiple players simultaneously, we propose a framework for extracting frequent patterns in multiple trajectories in football matches. As will be described in detail later, in the framework, we first convert a set of trajectories into a set of labeled sub-trajectories, or interval-based events. We then apply an extended version of frequent sequential pattern miner to the obtained events for extracting frequent combinations of interval-based events satisfying time and spatial constraints among the sub-trajectories. We introduce variables for subjects of each event into the pattern to distinguish the events played by the same player from those by difference ones. In addition, the variable introduction makes the miner possible to extract patterns by absorbing the difference of concrete player sets. The proposed enumeration algorithm can extract frequent patterns without duplication by employing a frequency count of a pattern in a long sequence as well as a representation of temporal relationship among interval-based events.

The proposed algorithm is evaluated quantitatively and qualitatively by using real trajectory datasets on nine matches in Japanese professional football league. Through the experiments, we confirm the effectiveness of the proposed algorithm from the aspects of the computation time and the number of extracted patterns. In addition, patterns for representing offensive and defensive tactical movements can be successfully extracted.

The rest of this paper is organized as follows: As preliminary, we give notation and definitions on trajectories and their derived events in section II. In section III, we define a frequent pattern to capture meaningful combinations of sub-trajectories by multiple subjects. An enumeration algorithm for the pattern is also proposed in this section. Experimental results using datasets on nine matches in Japanese professional football league are reported in section IV. Finally, in section V, we summarize the paper and describe future work.

II. Preliminaries

In this section, we give notation and definitions on a trajectory, which is the input of our pattern enumeration problem in this paper.

A. Trajectory database

A trajectory of length T by a subject o, represented as

$$tr^o = (o, \langle (x_1^o, y_1^o, a_1^o), \cdots, (x_t^o, y_t^o, a_t^o), \cdots, (x_T^o, y_T^o, a_T^o) \rangle\ ),$$

is a pair of o and a sequence of (x_t^o, y_t^o, a_t^o)s where (x_t^o, y_t^o) is the o's position in the 2-D space and a_t^o is a set of attributes of o at time $t\,(1 \leq t \leq T)$, respectively. In this paper, we assume that the positions and attributes can be obtained in constant time interval.

A sub-trajectory of tr^o from starting time e^+ to ending time e^- $(1 \leq e^+ \leq e^- \leq T)$ is a pair of the subject o and a part of o's trajectory from e^+ to e^-. It is defined formally as follows:

$$tr^o[e^+ : e^-] = (o, \langle (x_{e^+}^o, y_{e^+}^o, a_{e^+}^o), \cdots, (x_{e^-}^o, y_{e^-}^o, a_{e^-}^o) \rangle\ ).$$

A labeled sub-trajectory $ltr^o[e^+ : e^-]$ of a sub-trajectory $tr^o[e^+ : e^-]$ is defined as a five-tuple

$$\begin{aligned} &ltr^o[e^+ : e^-] = \\ &\quad (o, l, e^+, e^-, \langle (x_{e^+}^o, y_{e^+}^o, a_{e^+}^o), \cdots, (x_{e^-}^o, y_{e^-}^o, a_{e^-}^o) \rangle\ ) \end{aligned}$$

of subject o, label l, starting time e^+, ending time e^- and a part of trajectory from e^+ to e^-. In this paper, we assume

Y. Suzuki is with Graduate School of Integrated Basic Sciences, Nihon University, Japan.

T. Ozaki is with College of Humanities and Sciences, Nihon University, Japan. email : tozaki@chs.nihon-u.ac.jp

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

a labeling function or a model $\mathcal{M}$ which gives a label l for $tr^o[e^+ : e^-]$ based on a certain information on o and (x_t^o, y_t^o, a_t^o)s. Motif discovery algorithms[4], [5], [6], [7] and sub-trajectory clustering algorithms[8], [9], [10] are typical examples of such model. From technical reasons, a special label nil is introduced to show that the labeling model $\mathcal{M}$ determines that the given sub-trajectory $tr^o[e^+ : e^-]$ has no special information.

A set $TR = \{tr^{o_1}, \cdots, tr^{o_N}\}$ of N trajectories is called a trajectory database. By using a labeling function $\mathcal{M}$, a trajectory database TR can be converted into a set of labeled sub-trajectories. The obtained set is formally defined as follows:

$$\mathcal{M}(TR) = \left\{ ltr^o[e^+ : e^-] \;\middle|\; \begin{array}{l} o \in \{o_1, \cdots, o_N\} \\ l \neq nil \\ 1 \leq e^+ \leq e^- \leq T \end{array} \right\}.$$

We perform frequent pattern discovery on $\mathcal{M}(TR)$. The details will be described later in section III.

B. Constraints for a set of sub-trajectories

In this subsection, we introduce constraints on spatial spread and time difference for a set of labeled sub-trajectories. These constraints are utlized to define the support count of a pattern.

For a labeled sub-trajectory $ltr^o[e^+ : e^-]$, its center of gravity is defined as

$$\left(\overline{x^o_{e^+:e^-}}, \overline{y^o_{e^+:e^-}}\right) = \left(\frac{1}{L} \sum_{t=e^+}^{e^-} x_t^o, \; \frac{1}{L} \sum_{t=e^+}^{e^-} y_t^o \right)$$

where $L = e^- - e^+ + 1$. As similar, the center of gravity of a set $lTr = \{ltr^{o_1}[e_1^+ : e_1^-], \cdots, ltr^{o_n}[e_n^+ : e_n^-]\}$ of n labeled sub-trajectroies is defined as

$$\left(\overline{X_{lTr}}, \overline{Y_{lTr}}\right) = \left(\frac{1}{n} \sum_{i=1}^{n} \overline{x^{o_i}_{e_i^+:e_i^-}}, \; \frac{1}{n} \sum_{i=1}^{n} \overline{y^{o_i}_{e_i^+:e_i^-}} \right).$$

By using the above definitions, the degree d_1 of spatial spread of lTr is defined as a maximal distance between the center of gravity of each labeled sub-trajectory and that of lTr.

$$d_1(lTr) = \max_{ltr^o[e^+:e^-] \in lTr} \sqrt{\left(\overline{X_{lTr}} - \overline{x^o_{e^+:e^-}}\right)^2 + \left(\overline{Y_{lTr}} - \overline{y^o_{e^+:e^-}}\right)^2}$$

In addition, we consider another measure for the spatial spread without the center of gravities. Given two labeled sub-trajectories $ltr^{o_i}[e_i^+ : e_i^-]$ and $ltr^{o_j}[e_j^+, e_j^-]$ in lTr, the distance between the two is defined as the shortest distance between positions in $ltr^{o_i}[e_i^+ : e_i^-]$ and $ltr^{o_j}[e_j^+, e_j^-]$:

$$mdist\left(ltr^{o_i}[e_i^+ : e_i^-], ltr^{o_j}[e_j^+, e_j^-]\right) = \min_{e_i^+ \leq t_i \leq e_i^-, e_j^+ \leq t_j \leq e_j^-} \sqrt{(x^{o_i}_{t_i} - x^{o_j}_{t_j})^2 + (y^{o_i}_{t_i} - y^{o_j}_{t_j})^2} \; .$$

The maximum value of the shortest distances between any two labeled sub-trajectories in lTr is employed as a degree d_2 of spatial spread of lTr. It is formally defined as:

$$d_2(lTr) = \max_{x,y \in lTr} mdist(x, y).$$

Given a maximum threshold ρ, a degree $d_i\,(i \in \{1, 2\})$ of spatial spread, and a set of labeled sub-trajectories lTr, if the condition $d_i(lTr) \leq \rho$ holds, then we say that lTr satisfies the constraints on spatial spread and denote it as $s(lTr, d_i, \rho)$.

Given a set $lTr = \{ltr^{o_1}[e_1^+ : e_1^-], \cdots, ltr^{o_n}[e_n^+ : e_n^-]\}$ of n labeled subtrajectories, the starting and ending time of lTr is defined as

$$\begin{array}{rcl} start(lTr) & = & \min_{ltr^o[e^+:e^-] \in lTr} e^+ \\ end(lTr) & = & \max_{ltr^o[e^+:e^-] \in lTr} e^-, \end{array}$$

respectively. For a non-negative number τ and a set lTr, if lTr satisfies the condition $end(lTr) - start(lTr) \leq \tau$, denoted as $t(lTr, \tau)$, then lTr is said to satisfy the constraint on time difference.

C. Temporal relationship in a set of sub-trajectories

An event having time duration such as a labeled sub-trajectory is called an interval-based event in general. The temporal relationship among a set of interval-based events means the temporal relationship among endpoints. While a formal definition is provided later, the endpoint is a starting or ending point of events. In a naive way, the temporal relationship among n interval-based events can be represented by using ${}_{2n}C_2$ binary relationships on the all combination of endpoints. In contrast, Wu[11] proposes a sequence representation for temporal relationship among interval-based events using $2n$ endpoints and $2n - 1$ binary relationships only. We essentially utilize this nonambiguous representation to introduce a temporal sequence for a set of labeled sub-trajectories. The temporal sequence plays an essential role in the enumeration of frequent patterns.

The starting and ending points of a labeled sub-trajectory $(o, l, e^+, e^-, \langle \cdots \rangle)$ are denoted as $\langle o_i, l_i, e_i^+, e_i^- \rangle^+$ and $\langle o_i, l_i, e_i^+, e_i^- \rangle^-$. For a set lTr of n labeled sub-trajectories

$$lTr = \left\{ \left(o_1, l_1, e_1^+, e_1^-, \langle \cdots \rangle\right), \cdots, \left(o_n, l_n, e_n^+, e_n^-, \langle \cdots \rangle\right) \right\},$$

we introduce a set of endpoints

$$\begin{array}{r} EP(lTr) = \{\langle o_i, l_i, e_i^+, e_i^- \rangle^+ \mid \left(o_i, l_i, e_i^+, e_i^-, \langle \cdots \rangle\right) \in lTr\} \\ \cup \{\langle o_i, l_i, e_i^+, e_i^- \rangle^- \mid \left(o_i, l_i, e_i^+, e_i^-, \langle \cdots \rangle\right) \in lTr\} \end{array}$$

and four related functions $o(ep) = o$, $l(ep) = l$, $m(ep) = m$, and $t(ep) = e^{m(ep)}$ on $ep = \langle o, l, e^+ e^- \rangle^m \in EP(lTr)$. For example, given $ep = \langle o_i, l_i, e_i^+, e_i^- \rangle^+$, $m(ep)$ becomes $+$ and thus $t(ep)$ becomes $e_i^{m(ep)} = e_i^+$.

Under the above preparation, given a set lTr, a sequence

$$seq(lTr) = \langle ep_1, ep_2, \cdots, ep_{2n-1}, ep_{2n} \rangle$$

satisfying the condition

$$\{ep_1, \cdots, ep_{2n}\} = EP(lTr) \wedge \forall i, j_{(1 \leq i < j \leq 2n)} \, [ep_i \prec ep_j]$$

is called a temporal sequence of lTr. A binary relationship $\prec$ between two endpoints ep_i and ep_j becomes true if one of the following condition holds:

1) $t(ep_i) < t(ep_j)$
2) $t(ep_i) = t(ep_j) \wedge m(ep_i) = + \wedge m(ep_j) = -$
3) $t(ep_i) = t(ep_j) \wedge m(ep_i) = m(ep_j) \wedge$ $o(ep_i)$ alphabetically precedes $o(ep_j)$

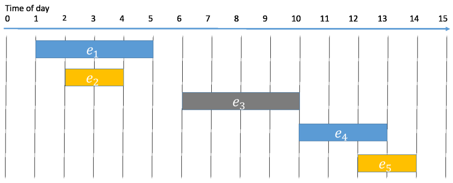

Fig. 1. An example of sequence of labeled sub-trajectories

4) $t(ep_i) = t(ep_j) \wedge m(ep_i) = m(ep_j) \wedge o(ep_i) = o(ep_j)$ $l(ep_i)$ alphabetically precedes $l(ep_j)$

This representation can be regarded as a natural extension of that by [11] to handle the subjects of events. Therefore, while the proof is not provided in this paper, the temporal sequence of a set of labeled sub-trajectories can be determined uniquely. Please refer to the original paper[11] for details.

We show an example of labeled sub-trajectories and its temporal sequence. Given a set of five labeled sub-trajectories

$$lTr = \left\{ \begin{array}{ll} (o_1, y, 1, 5, \langle\cdots\rangle), & (o_2, b, 2, 4, \langle\cdots\rangle), \\ (o_3, r, 6, 10, \langle\cdots\rangle), & (o_4, y, 10, 13, \langle\cdots\rangle), \\ (o_5, b, 12, 14, \langle\cdots\rangle) & \end{array} \right\}$$

shown in Fig.1, a temporal sequence of lTr becomes

$$\begin{array}{l} \langle \langle o_1, y, 1, 5\rangle^+,\ \langle o_2, b, 2, 4\rangle^+,\ \langle o_2, b, 2, 4\rangle^-,\ \langle o_1, y, 1, 5\rangle^-, \\ \langle o_3, r, 6, 10\rangle^+,\ \langle o_4, y, 10, 13\rangle^+,\ \langle o_3, r, 6, 10\rangle^-, \\ \langle o_5, b, 12, 14\rangle^+,\ \langle o_4, y, 10, 13\rangle^-,\ \langle o_5, b, 12, 14\rangle^- \rangle. \end{array}$$

III. Frequent patterns in trajectory databases

In this section, we give a formal definition of our patterns and data mining problem. We also propose an enumeration algorithm for the problem.

A. Pattern

An abstracted labeled sub-trajectory (O, l, E^+, E^-) consists of a concrete label l and three variables O for subject, E^+ for starting time and E^- for ending time. It represents a proposition "the label of O's sub-trajectory from E^+ to E^- is l". We use the term "abstracted" since an abstracted labeled sub-trajectory can be obtained from a labeled sub-trajectory by introducing variables and ignoring the concrete sub-trajectory.

Given a set of n abstracted sub-trajectories $alTr$, its set of all endpoints is denoted as

$$aEP(alTr) = \bigcup_{(O,l,E^+,E^-)\in alTr} \left\{ \langle O, l, E^+, E^-\rangle^+, \langle O, l, E^+, E^-\rangle^- \right\}.$$

A sequence $\langle EP_1, \cdots, EP_{2n}\rangle$ is called a temporal sequence of $alTr$, if $\{EP_1, \cdots, EP_{2n}\} = aEP(alTr)$ holds. We interpret the condition $\forall i_{1\le i\le 2n-1}\,[EP_i \prec EP_{i+1}]$ holds in the sequence. Note that, different abstracted labeled sub-trajectories and their endpoints use same variables in common to represent the same subjects, starting times or ending times. In addition, all variables are required to satisfy the object identity assumption[12], and thus the different variables represent different instances.

In this paper, the temporal sequence of abstracted labeled sub-trajectories is employed as a pattern language, since it captures a combination of labeled sub-trajectories by (multiple) subject(s) with temporal relationship. Note that, a combination of (abstracted) labeled sub-trajectories can be obtained by its temporal sequence. We show an example of pattern below. A temporal sequence

$$\begin{array}{l} \langle\quad \langle O_a, r, E_a, E_b\rangle^+,\ \langle O_b, y, E_b, E_c\rangle^+, \\ \quad\ \langle O_a, r, E_a, E_b\rangle^-,\ \langle O_b, y, E_b, E_c\rangle^- \quad \rangle \end{array}$$

is a pattern which represents that: a subject O_a starts an event r at time E_a and ends it at time E_b. At the same time O_a finishes r, a subject O_b starts an event y and finishes it at E_c.

We introduce a general-to-specific relationship among patterns. Given two patterns α and β, if there exists a variable to variable substitution θ under the object identity assumption such that $\alpha\theta \sqsubseteq \beta$, then we say that α is more general than β, and denote it $\alpha \preceq \beta$. Conversely, β is said to be more specific than α. The binary relationship $x \sqsubseteq y$ means that all elements in x appear in y in the same order they appear in x.

For example, a pattern

$$\begin{array}{l} \alpha = \langle\quad \langle O_a, r, E_a, E_b\rangle^+,\ \langle O_b, y, E_b, E_c\rangle^+, \\ \qquad\quad \langle O_a, r, E_a, E_b\rangle^-,\ \langle O_b, y, E_b, E_c\rangle^- \quad \rangle \end{array}$$

is more general than a pattern

$$\begin{array}{l} \beta = \langle\langle O_a, r, E_a, E_b\rangle^+,\ \langle O_b, y, E_b, E_c\rangle^+, \langle O_a, r, E_a, E_b\rangle^- \\ \qquad \langle O_c, b, E_d, E_e\rangle^+, \langle O_b, y, E_b, E_c\rangle^-, \langle O_c, b, E_d, E_e\rangle^-\rangle. \end{array}$$

B. Occurrence and frequency of patterns

Given a trajectory database TR and a pattern p, a set $lTr \subseteq \mathcal{M}(TR)$ of label sub-trajectories is said to be an occurrence of p in TR, if there exists an substitution θ of variables such that $seq(lTr) = p\theta$.

For example, a set of two labeled sub-trajectories

$$occ = \{(o_3, r, 6, 10, \langle\cdots\rangle), (o_4, y, 10, 13, \langle\cdots\rangle)\}$$

is an occurrence of a pattern

$$\begin{array}{l} p = \langle\quad \langle O_a, r, E_a, E_b\rangle^+,\ \langle O_b, y, E_b, E_c\rangle^+, \\ \qquad\quad \langle O_a, r, E_a, E_b\rangle^-,\ \langle O_b, y, E_b, E_c\rangle^- \quad \rangle \end{array}$$

since a substitution $\theta = \{o_3/O_a, o_4/O_b, 6/E_a, 10/E_b, 13/E_c\}$ makes $p\theta$ be identical with

$$\begin{array}{l} seq(occ) = \langle\quad \langle o_3, r, 6, 10\rangle^+,\ \langle o_4, y, 10, 13\rangle^+, \\ \qquad\qquad\quad \langle o_3, r, 6, 10\rangle^-,\ \langle o_4, y, 10, 13\rangle^- \quad \rangle. \end{array}$$

A set of all occurrences of a pattern p in TR is denoted as

$$Occ(TR, p) = \{lTr \subseteq \mathcal{M}(TR) \mid \exists\theta\ s.t.\ seq(lTr) = p\theta\}.$$

We consider the occurrences satisfying the constraints on spatial spread and time difference. Given a maximum threshold ρ and the degree d_i $(i \in \{1, 2\})$ of spatial spread, a set of all occurrences satisfying the spatial constraint is denoted as

$$Occ^{\rho}(TR, P) = \{lTr \in Occ(TR, p) \mid s(lTr, d_i, \rho)\}.$$

As similar,

$$Occ_{\tau}(TR, P) = \{lTr \in Occ(TR, p) \mid t(lTr, \tau)\}$$

denotes a set of all occurrences satisfying the constraint on time difference τ. By using the two sets, a set of all occurrences satisfying both constraints is defined as

$$Occ^{\rho}_{\tau}(TR, P) = Occ^{\rho}(TR, P) \cap Occ_{\tau}(TR, P).$$

Several frequency measures for sequential patterns in a long single sequence of point-based events have been proposed[13], [14], [15] in the past. Among them, we employ the measure in [13], and apply it to define the frequency of our patterns using a set of occurrences.

Given a pattern p and a trajectory database TR, the unconditional head frequency of p in TR is defined as

$$H\_Freq(TR, p) = |\ \{start(lTr) \mid lTr \in Occ(TR, p)\}\ |.$$

In other words, we regard the number of starting times the set of labeled sub-trajectories begins as a frequency. As similar, head frequency under spatial and temporal constraints can be defined as follows:

$$\begin{aligned} H\_Freq^{\rho}(TR, p) &= |\{start(lTr) \mid lTr \in Occ^{\rho}(TR, p)\}| \\ H\_Freq_{\tau}(TR, p) &= |\{start(lTr) \mid lTr \in Occ_{\tau}(TR, p)\}| \\ H\_Freq^{\rho}_{\tau}(TR, p) &= |\{start(lTr) \mid lTr \in Occ^{\rho}_{\tau}(TR, p)\}| \end{aligned}$$

While H_Freq^{ρ} and H_Freq_{τ} are frequencies under the spatial and temporal constraint respectively, we consider both constraints in $H\_Freq^{\rho}_{\tau}$.

The total frequency of a pattern p is defined as the minimum value of the head frequency of arbitrary patterns which are more general than p. We consider four total frequencies based on the combinations of spatial and temporal constraints. They are formally defined as follows:

$$\begin{cases} T\_Freq(TR, p) = \min_{\gamma \preceq \alpha} (\, H\_Freq(TR, p)\,) \\ T\_Freq^{\rho}(TR, p) = \min_{\gamma \preceq \alpha} (\, H\_Freq^{\rho}(TR, p)\,) \\ T\_Freq_{\tau}(TR, p) = \min_{\gamma \preceq \alpha} (\, H\_Freq_{\tau}(TR, p)\,) \\ T\_Freq^{\rho}_{\tau}(TR, p) = \min_{\gamma \preceq \alpha} (\, H\_Freq^{\rho}_{\tau}(TR, p)\,) \end{cases}$$

Since a set $Occ^{\rho}_{\tau}(TR, P)$ is a subset of $Occ_{\tau}(TR, P)$, $H\_Freq^{\rho}_{\tau}(TR, P) \leq H\_Freq_{\tau}(TR, P)$ must hold, and it implies the relation $T\_Freq^{\rho}_{\tau}(TR, P) \leq T\_Freq_{\tau}(TR, P)$.

Note that, while the proof is not provided, the frequency T_Freq_{τ} has the anti-monotone property on the pattern's generality and the following relatioship holds:

$$\forall \alpha, \beta\, [\alpha \preceq \beta \rightarrow T\_Freq_{\tau}(TR, \alpha) \geq T\_Freq_{\tau}(TR, \beta)].$$

However, the frequencies T_Freq^{ρ} and $T\_Freq^{\rho}_{\tau}$ do not have such property unfortunately in general.

Here, we define our data mining problem in this paper. Given a trajectory database TR, a labeling model $\mathcal{M}$, a maximum threshold ρ for the spatial spread, a maximum threshold τ for the time difference, and a minimum threshold σ for the frequency, our data mining problem is to enumerate all patterns p satisfying the condition $T\_Freq^{\rho}_{\tau}(TR, p) \geq \sigma$.

C. Enumeration algorithm

To enumerate all frequent patterns, we employ a breadth first search strategy since the total frequency $T\_Freq^{\rho}_{\tau}$ of a pattern α requires all head frequencies $T_Freq^{\rho}_{\tau}$s of γ such that $\gamma \preceq \alpha$. In addition, we utilize the idea of the reverse search[16] to constract a tree-shaped pattern space.

The last element $EP_i = \langle O_i, l_i, E^{+}_i, E^{-}_i \rangle^{+}$ in a pattern p such that $m(EP_i) = +$ is denoted as last(p,+). In addition, last(p,-) denotes a element $\langle O_i, l_i, E^{+}_i, E^{-}_i \rangle^{-}$ in p which is another endpoint corresponding to last(p,+). If two patterns $\alpha = \langle EP^{\alpha}_1, \cdots, EP^{\alpha}_{2n} \rangle$ and $\beta = \langle EP^{\beta}_1, \cdots, EP^{\beta}_{2(n+1)} \rangle$ satisfy the condition $\alpha = \beta \setminus \langle last(\beta, +), last(\beta, -) \rangle$ holds, then α is called a parent of β, and β is said to be a child of α conversely where $\setminus$ denotes a subtraction of elements.

```
TP_LS^TR($(TR, \sigma, \tau, \rho)$)
    Input
      $TR$ : a trajectory database
      $\sigma$ : minimum support threshold
      $\tau$ : maximum threshold on time difference
      $\rho$ : maximum threshold on spatial spread
    Output
      $\mathcal{F}$ : a set of frequent patterns
    $P_1 \leftarrow$set of frequent events
    $\mathcal{F} \leftarrow P_1$
    TP_LS^TR($(TR, \sigma, \tau, \rho, P_1, \mathcal{F})$)
    return $\mathcal{F}$

TP_LS^TR($(TR, \sigma, \tau, \rho, C, \mathcal{F})$)
    if $C = \emptyset$ then return
    $C' \leftarrow \{\}$
    for each $p \in C$
      for each $p' \in extension(p)$
        if $T\_Freq_{\tau}(TR, p') \geq \sigma$
          then $C' \leftarrow C' \cup \{p'\}$
        if $T\_Freq^{\rho}_{\tau}(TR, p') \geq \sigma$
          then $\mathcal{F} \leftarrow \mathcal{F} \cup \{p'\}$
    TP_LS^TR($(TR, \sigma, \tau, \rho, C', \mathcal{F})$)
```

Fig. 2. The algorithm $\mathrm{TP\_LS^{TR}}$ for the enumeration of frequent patterns

We can obtain a child of a pattern α using the parent-children relationship inversely. A child γ of a pattern α can be derived by inserting two endpoints of one abstracted labeled sub-trajectory into any position after last(α,+). At this time, we consider all combinations of variables for the subject, starting and ending time from a set of existing variables in α and new variables. Note that, this extension can be regarded as an extension of the expansion strategy in the TPrefixSpan algorithm[11] for enumerating sequential patterns in a long single interval-based events.

The proposed algorithm named $\mathrm{TP\_LS^{TR}}$ is shown in Fig.2. In the algorithm, "extension" returns a set of children of given pattern p. While the value of T_Freq_{τ} is utilized as a pruning criterion because of its anti-monotone property, the value of $T\_Freq^{\rho}_{\tau}$ is just used for determining whether or not the pattern is frequent. The derived frequent patterns are stored in $\mathcal{F}$.

IV. Experiments and consideration

A. Data set

In order to evaluate the effectiveness of the proposed framework, we implement our algorithm $\mathrm{TP\_LS^{TR}}$ using Java language and conduct experiments using nine matches in the J. League (Japanese professional football league) held in 2015 and 2016. The original dataset is provided by Data Stadium Inc. [1] , and we derive 207 (= $(1 + 22) \times 9$) of trajectories of ball and players in every 0.2 second interval.

[1] https://www.datastadium.co.jp

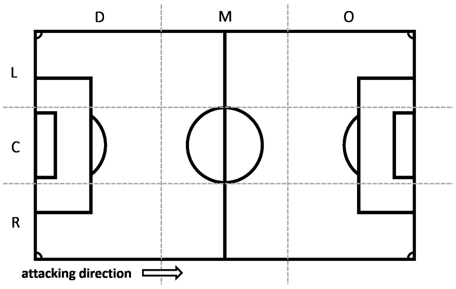

Fig. 3. Area on the field

TABLE I
NUMBER OF LABELS AND LABELED SUB-TRAJECTORIES

time width	3	4	5
# of labeled sub-trajectories	429,169	387,513	218,485
# of labels	13	15	23

In the experiments, each trajectory is divided into sub-trajectories with a time width of $\{3, 4, 5\}$ seconds and a slide width of 2 seconds. After the division, the COIN clustering approach[4] is applied to extract meaningful movements or motifs. A labeling function $\mathcal{M}$ is built as a combination of the id of motifs and that of the area in the field shown in Fig.3 where the (sub-) trajectory is observed. We prepare nine areas by dividing the whole football field vertically and horizontally into three, respectively. In the vertical direction, we prepare three labels "D" for the defensive area, "M" for the area near from the center circle, and "O" for the attacking area. On the other hand, in the horizontal direction, three labels "L", "C" and "R" represent the left, center and right area, respectively. The combination of vertical and horizontal labeled are used to represent the area. For example, the leftmost and uppermost area is denoted as "DL".

The number of obtained labels and labeled sub-trajectories are summarized in TABLE I.

B. Quantitative evaluation

We measure the number of obtained patterns and execution times by using a Linux machine (CPU: Intel Xeon 3.20 GHz, main memory: 24 GB). In the experiments, the minimum frequency threshold σ, the maximum time difference τ, and the maximum spatial spread ρ are set to $\sigma \in \{100, 150, 200\}$, $\tau \in \{10, 15, 20\}$ seconds and $\rho \in \{20m\ in\ d_1, 10m\ in\ d_2\}$, respectively. In addition, we restrict that the patterns must have less than or equal to five abstracted labeled sub-trajectories. The results are summarized in TABLE II.

TABLE II
THE NUMBER OF OBTAINED PATTERNS AND EXECUTION TIME

	number of patterns			execution time(minuts)		
$k \backslash \sigma$	100	150	20	100	150	200
d_1, $s = 20m$						
10	3,844	2,309	729	39	33	9
15	72,411	44,019	25,538	152	116	83
20	667,943	509,126	22,7310	529	354	264
d_2, $s = 10m$						
10	4,571	2,951	837	57	37	22
15	76,985	56,381	31,388	438	232	152
20	724,605	591,723	282,654	699	556	413

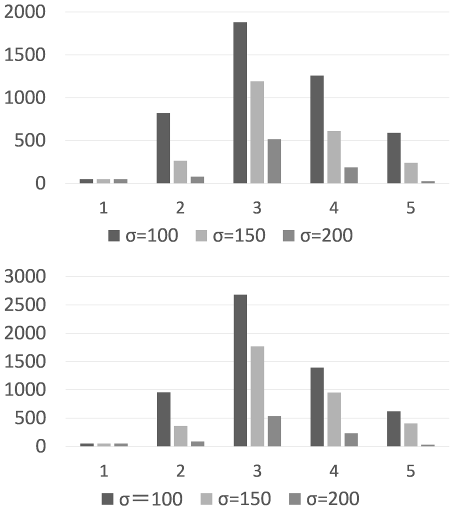

Fig. 4. Distribution of the number of abstracted labeled sub-trajectories in a pattern (top: d_1, $k = 10$, bottom: d_2, $k = 10$)

From the results, regardless of the parameter settings, we can confirm that the proposed algorithm achieved the pattern enumeration within a reasonable computation time. In addition, we recognize a similar tendency between d_1 and d_2 with respect to the change in the number of extracted patterns for parameters σ and τ.

The distributions of the numbers of abstracted labeled sub-trajectories in a pattern are shown in Fig.4 under the settings of $\tau = 10$ and $\sigma \in \{100, 150, 200\}$. As the results show, it can be seen that many patterns having three and four abstracted labeled sub-trajectories are obtained regardless of the value of σ and the definition d_i for the spatial spread. This result seems to represent the diversity of actions and relationships among players in football. In addition, we believe that many basic but important relationships such as "third man running" are extracted.

C. Qualitative evaluation

Through the experiments, we successfully obtained two patterns which capture certain defensive and offensive tactical movements. The patterns are shown below:

$$
\begin{aligned}
p_D = \langle\ & \langle A, 1{:}CD, E_1, E_2\rangle^+, \langle B, 3{:}CD, E_1, E_2\rangle^+, \\
& \langle C, 1{:}LD, E_4, E_5\rangle^+, \langle A, 1{:}CD, E_1, E_2\rangle^-, \\
& \langle B, 3{:}CD, E_1, E_2\rangle -, \langle C, 1{:}LD, E_4, E_5\rangle^- \quad \rangle \\
p_O = \langle\ & \langle A, 12{:}LM, E_1, E_2\rangle^+, \langle B, 9{:}LM, E_3, E_4\rangle^+, \\
& \langle C, 9{:}LM, E_5, E_6\rangle^+, \langle A, 12{:}LM, E_1, E_2\rangle^-, \\
& \langle B, 9{:}LM, E_3, E_4\rangle^-, \langle C, 9{:}LM, E_5, E_6\rangle^- \quad \rangle
\end{aligned}
$$

In the patterns, each label $M{:}A$ represents that a movement, or motif M, is performed in the area A.

An example of the occurrence of the pattern is shown in Fig.5 and Fig.6, respectively. The pattern p_D can be interpreted as that three defensive players A, B and C work together in their own team territory and move forward to raise the defensive line. On the other hand, we can observe in p_O that a player A overlaps from own team territory ($12{:}LM$) and other player B moves toward center ($9{:}LM$) to open the space for the player A.

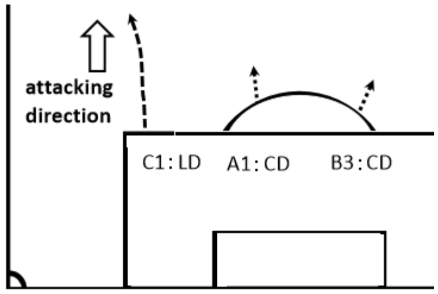

Fig. 5. An occurrence of p_D

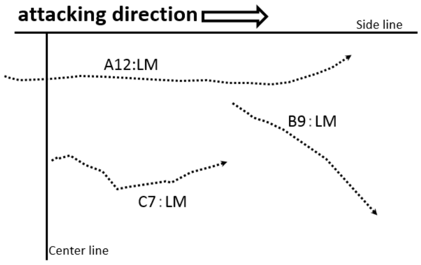

Fig. 6. An occurrence of p_O

V. Conclusion

In this paper, we proposed a frequent pattern enumeration algorithm in a trajectory database as a first step towards the extraction of meaningful tactical movements in football matches. The algorithm first converts the trajectory database into a set of labeled sub-trajectories, and then performs the enumeration based on the pattern expansion for the interval-based events and the frequency count with the consideration of spatial and temporal constraints. The effectiveness of the proposed algorithm was evaluated quantitatively and qualitatively using trajectories in nine real professional football matches.

The proposed algorithm requires the parameters on time and spatial spread. However, the tactical behaviors in football have various temporal and spatial scales, and it is difficult to set the parameters appropriately in advance. Therefore, as one of our future work, it is necessary to prepare a certain optimization algorithm for the parameter. In addition, due to the characteristics of the frequent pattern discovery, many similar patterns can be extracted. It is also an important research direction to evaluate the obtained patterns quantitatively and to perform pattern filtering by statistical significance.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number 17K00315.

References

[1] C.-H. Kang, J.-R. Hwang, and K.-J. Li, “Trajectory analysis for soccer players,” in *Proc. of the 6th IEEE International Conference on Data Mining Workshops*. IEEE, 2006, pp. 377–381.

[2] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. Matthews, “Assessing team strategy using spatiotemporal data,” in *Proc. of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2013, pp. 1366–1374.

[3] A. Bialkowski, P. Lucey, P. Carr, I. Matthews, S. Sridharan, and C. Fookes, “Discovering team structures in soccer from spatiotemporal data,” *IEEE Transactions on Knowledge Data Engineering*, vol. 28, no. 10, pp. 2596–2605, 2016.

[4] P. Agarwal, G. Shroff, S. Saikia, and Z. Khan, “Efficiently discovering frequent motifs in large-scale sensor data,” in *Proc. of the 2nd ACM IKDD Conference on Data Sciences*, ACM, 2015, pp. 98–103.

[5] B. Tang, M. L. Yiu, K. Mouratidis, and K. Wang, “Efficient motif discovery in spatial trajectories using discrete fréchet distance,” in *Proc. of the 20th International Conference on Extending Database Technology*, OpenProceedings.org, 2017, pp. 378–389.

[6] T. Oates, A. P. Boedihardjo, J. Lin, C. Chen, S. Frankenstein, and S. Gandhi, “Motif discovery in spatial trajectories using grammar inference,” in *Proc. of the 22nd ACM International Conference on Information & Knowledge Management*, ACM, 2013, pp. 1465–1468.

[7] J. Grabocka, N. Schilling, and L. Schmidt-Thieme, “Latent time-series motifs,” *ACM Transactions on Knowledge Discovery from Data*, vol. 11, no. 1, pp. 6:1–6:20, Jul. 2016.

[8] G. Yang, Z. Huang, and X. Wang, “Comparison study of sub-trajectory clustering in data mining,” *IOP Conference Series: Earth and Environmental Science*, vol. 69, no. 1, p. 012143, 2017.

[9] J.-G. Lee, J. Han, and K.-Y. Whang, “Trajectory clustering: A partition-and-group framework,” in *Proc. of the 2007 ACM SIGMOD International Conference on Management of Data*, ACM, 2007, pp. 593–604.

[10] C. Chang and B. Zhou, “Multi-granularity visualization of trajectory clusters using sub-trajectory clustering,” in *Proc. of 9th IEEE International Conference on Data Mining Workshops*, IEEE, 2009, pp. 577–582.

[11] S.-Y. Wu and Y.-L. Chen, “Mining nonambiguous temporal patterns for interval-based events,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 6, pp. 742–758, 2007.

[12] G. Semeraro, F. Esposito, D. Malerba, N. Fanizzi, and S. Ferilli, “A logic framework for the incremental inductive synthesis of datalog theories,” in *Proc. of the 7th International Workshop on Logic Programming Synthesis and Transformation*, 1997, pp. 300–321.

[13] K. Iwanuma, R. Ishihara, Y. Takano, and H. Nabeshima, “Extracting frequent subsequences from a single long data sequence: a novel anti-monotonic measure and a simple on-line algorithm,” in *Proc. of the 5th IEEE International Conference on Data Mining*. IEEE, 2015, pp. 186–193.

[14] A. Achar, S. Laxman, and P. S. Sastry, “A unified view of the apriori-based algorithms for frequent episode discovery,” *Knowledge and Information Systems*, vol. 31, no. 2, pp. 223–250, May 2012.

[15] J. Yang, W. Wang, and P. S. Yu, “Mining surprising periodic patterns,” *Data Mining and Knowledge Discovery*, vol. 9, no. 2, pp. 189–216, Sep 2004.

[16] D. Avis and K. Fukuda, “Reverse search for enumeration,” *Discrete Applied Mathematics*, vol. 65, no. 1, pp. 21 – 46, 1996,