SCIENTIFIC REPORTS
nature research

OPEN

Clustering algorithm for formations in football games

Takuma Narizuka[1] & Yoshihiro Yamazaki[2]

Received: 2 April 2019
Accepted: 6 August 2019
Published online: 11 September 2019

In competitive team sports, players maintain a certain formation during a game to achieve effective attacks and defenses. For the quantitative game analysis and assessment of team styles, we need a general framework that can characterize such formation structures dynamically. This paper develops a clustering algorithm for formations of multiple football (soccer) games based on the Delaunay method, which defines the formation of a team as an adjacency matrix of Delaunay triangulation. We first show that heat maps of entire football games can be clustered into several average formations: "442", "4141", "433", "541", and "343". Then, using hierarchical clustering, each average formation is further divided into more specific patterns (clusters) in which the configurations of players are different. Our method enables the visualization, quantitative comparison, and time-series analysis for formations in different time scales by focusing on transitions between clusters at each hierarchy. In particular, we can extract team styles from multiple games regarding the positional exchange of players within the formations. Applying our algorithm to the datasets comprising football games, we extract typical transition patterns of the formation for a particular team.

In competitive team sports, such as football (soccer) and basketball, each player coordinates with team members and interacts with opposing players. Throughout such interactions, players maintain a certain formation at the team level. Such a formation structure reflects a team's strategies for achieving effective attacks and defenses in order to win[1–4]. A traditional method of characterizing formations employs notation such as "4-4-2", which indicates four defenders, four midfielders, and two forwards. Although this is a convenient means of roughly grasping formation structures, such static notation is too simple to analyze real games. In fact, the following more quantitative methods have been introduced.

The first example is based on a Voronoi region defined for each player, which is the set of field locations whose distances from the player are less than from any other[5]. Intuitively, this corresponds to the territory of the player on the field. The basic properties of the Voronoi region have been investigated for football and hockey games[6,7], and modified version considering the velocity and acceleration of a player have also been proposed[8–12].

Bialkowski *et al*. developed another approach to formations, called "role representation"[13,14]. Here, the "role" represents the relative position of each player in the formation such as "center forward" or "left wing". The key idea behind the role representation is that players are not distinguished by their identities such as uniform numbers or their names, but rather by the role numbers assigned to them; the formation of a team is defined as the set of roles. Although the player identity is fixed throughout a game, the role can change during a game depending on their relative positions. While the previous notation such as "4-4-2" is static, the role representation enables more dynamical characterization of formations, e.g., exchange of players' roles during a game.

Along with these studies, we have proposed the Delaunay method, which identifies the adjacency relationships of players' Voronoi regions, i.e., the Delaunay network, with the formation of a team[15] (see Methods for details). Because the formation at time t is quantified by an adjacency matrix $A(t)$ in this method, dissimilarity measures between two different formations can be defined. On the basis of the Delaunay method, we have also proposed a clustering algorithm for formations in a single game. This algorithm divides Delaunay networks, which are given at every unit time in a single game, into clusters by means of hierarchical clustering. We have demonstrated that our method can characterize the differences and dynamics of football formations at different time resolutions within a game by controlling the number of clusters.

The Delaunay method is useful for the quantitative comparison and time-series analysis of formations. However, comparison of formations among different games is not available at present, because the above clustering algorithm for a single game cannot be straightforwardly extended to the case of multiple games. The problem

[1]Department of Physics, Faculty of Science and Engineering, Chuo University, Bunkyo, Tokyo, 112-8551, Japan. [2]Department of Physics, School of Advanced Science and Engineering, Waseda University, Shinjuku, Tokyo, 169-8555, Japan. Correspondence and requests for materials should be addressed to T.N. (email: pararel@gmail.com)

is as follows. In order to quantify a formation using an adjacency matrix $A(t)$, uniform numbers $\overrightarrow{U} = [a, b, \ldots, j]$ of players (player identities) need to be assigned to the indexes $\overrightarrow{I} = [1, 2, \ldots, 10]$ of $A(t)$. If we cluster Delaunay networks of a single game in which no player substitutions occur, then an arbitrary correspondence between $\overrightarrow{U}$ and $\overrightarrow{I}$ can be employed. However, clustering over multiple games requires the assignment of multiple uniform numbers $\overrightarrow{U_1}$, $\overrightarrow{U_2}$, ... for different games to one set of indexes $\overrightarrow{I}$, and such an assignment is not uniquely determined.

For the application of the Delaunay method to the real game analysis, we have to deal with this problem. In fact, the difference of formations among multiple games is essential information for the assessment of teams' styles or strategies. In this paper, we propose an extended algorithm that can cluster formations over multiple games, and demonstrate the formation analysis by applying our algorithm to the datasets comprising football games of Japan professional football league (J League). Our method first clusters heat maps in multiple football games into several average formations: "442", "4141", "433", "541", and "343". Then, we employ the role representation and hierarchical clustering, and each average formation is further divided into more specific patterns in which the configurations of players are slightly different. Based on the transition network between clusters, we extract typical transition patterns of the formation for a particular team.

Methods

Dataset and analysis. We employ datasets comprising 45 football games by 18 teams of the top league of J League (J1 League) second stage 2016, provided by DataStadium Inc., Japan. The DataStaduim has been authorized to collect and sell data under a contract with the J League. This contract also ensures that the use of relevant datasets does not infringe any rights of players and clubs belonging to J League. The datasets are not open. We have received permission to use them for this research from the DataStadium. The list of names of 18 teams is as follows:

"Fukuoka", "Hiroshima", "Iwata", "Kashima", "Kashiwa", "Kawasaki",
"Kobe", "Kofu", "Nagoya", "Niigata", "Omiya", "Osaka",
"Sendai", "Shonan", "Tokyo", "Tosu", "Urawa", "Yokohama".

There are five games per team, and each of five games was taken place on Sept. 25, Oct. 1, Oct. 22, Oct. 29, and Nov. 3 in 2016. In this paper, we refer to each game in a form such as Sept. 25 game of "Sendai". Each dataset contains all players' absolute positions every 0.04 seconds (i.e., the frame rate is 25 fps), which are tracked automatically by multiple cameras fixed in each stadium; the spatial resolution of the data is centimeter scale. For simplicity, we focus on the 10 players ($N = 10$) other than the goalkeeper for each team and analyze the data of the first halves of games. It is noted that we exclude several games with player substitutions in the first half from the analysis.

Our analysis is performed using python packages; for the calculation of Voronoi region and Delaunay triangulation, *Voronoi* and *Delaunay* classes in the *scipy.spatial* module was used; for the hierarchical clustering, *linkage* class in the *scipy.cluster.hierarchy* module was used. All calculations were executed on a MacBook Pro with a 2 GHz Intel Core i5 processor and 16 GB of memory.

The absolute coordinates of the j-th player of a team at time t is denoted as $\overrightarrow{R_j}(t)$. The centroid position and standard deviation of a team respectively defined as follows:

$$\overrightarrow{r_c}(t) = \frac{1}{N}\sum_{j=1}^{N} \overrightarrow{r_j}(t), \tag{1}$$

$$\sigma(t) = \sqrt{\frac{1}{N}\sum_{j=1}^{N} |\overrightarrow{r_c}(t) - \overrightarrow{r_j}(t)|^2}\,. \tag{2}$$

Using $\overrightarrow{r_c}(t)$ and $\sigma(t)$, the normalized coordinates $\overrightarrow{r_j}(t)$ for the j-th player are calculated as

$$\overrightarrow{R_j}(t) = \frac{\overrightarrow{r_j}(t) - \overrightarrow{r_c}(t)}{\sigma(t)}. \tag{3}$$

Delaunay method and clustering algorithm for a single game. Here, we summarize the Delaunay method and the clustering algorithm for a single team[15]. In our method, we regard a football formation as adjacency relationships of players, which is independent of the deviation $\sigma(t)$ of the team. Specifically, as shown in Fig. 1(a), a formation of a team at time t is quantified using the adjacency matrix $A(t)$ of the Delaunay network, whose components $A_{ij}(t)$ are given by

$$A_{ij}(t) = \begin{cases} 1 & \text{if the Voronoi regions of players } i \text{ and } j \text{ are adjacent with each other at } t, \\ 0 & \text{otherwise.} \end{cases}$$

Although there are other options for the definition of neighbors in 2D space, we choose the Delaunay triangulation because it is reasonable for the visualization and clustering of formations as shown below.

Owing to this quantification, a dissimilarity measure between two formations at different times can be introduced as

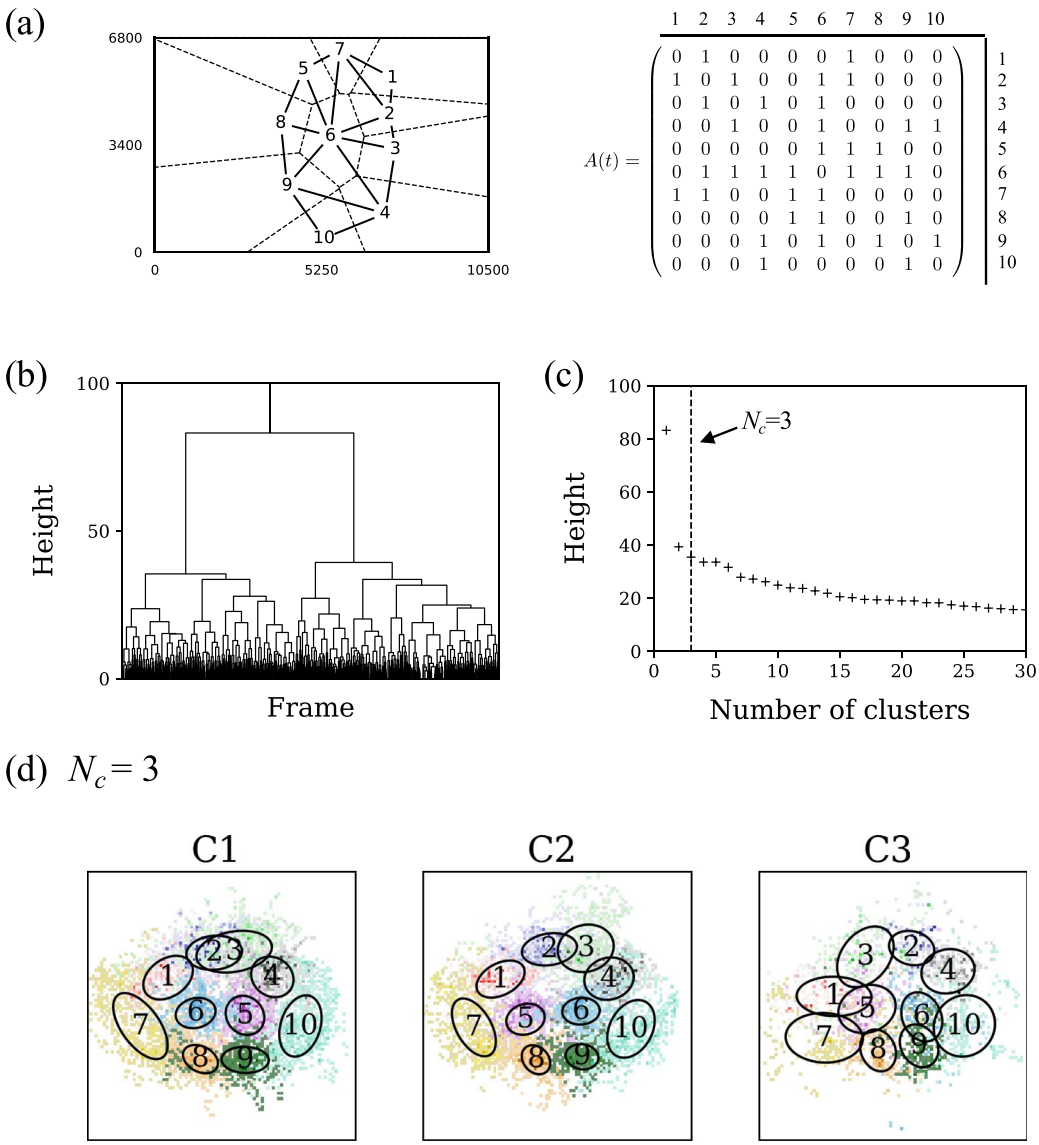

Figure 1. A typical example of a clustering process for Sept. 25 game of "Sendai". (**a**) A Delaunay network at a certain frame and its adjacency matrix. The unit of the horizontal and vertical axes is centimeter. (**b**) The Dendrogram and (**c**) the relation between the number of clusters and height, which are obtained from the hierarchical clustering. The height of the vertical axes corresponds to the distance between two merged clusters. (**d**) Coarse-grained formations for $N_c = 3$ in the normalized coordinates where the direction of offense is upward. The major difference between clusters is that players 2 and 3, and players 5 and 6 exchange their positions.

$$D_{tt'} = \|A(t) - A(t')\|^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}[A_{ij}(t) - A_{ij}(t')]^2. \tag{4}$$

Here, we define $D_{tt'}$ as the Euclidean squared distance, considering the hierarchical clustering using Wards' method. The dissimilarity $D_{tt'}$ becomes large when a number of edges are rewired due to the positional exchange of players within the formations.

Based on this dissimilarity measure, we introduced a clustering algorithm for formations appearing in a single game through the following four steps (i)-(iv). (i) The Delaunay networks every Δf frames in a single game are computed, where the frame rate is 25 fps. (ii) Hierarchical clustering is performed using Ward's method[16], where the input to the clustering is the dissimilarity matrix D whose components are $D_{tt'}$ defined by Eq. (4). In the Wards' method, distance between two clusters C_1 and C_2 is given by

$$h(C_1, C_2) = \frac{2n_1n_2}{n_1 + n_2}\left\|\frac{1}{n_1}\sum_{t_1\in C_1}A(t_1) - \frac{1}{n_2}\sum_{t_2\in C_2}A(t_2)\right\|^2, \tag{5}$$

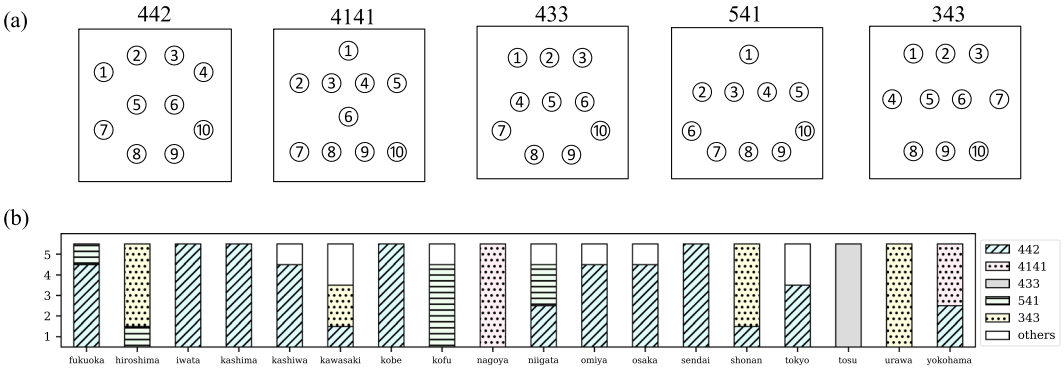

Figure 2. (**a**) Schematic representation of each average formation. The heat map of each team shown in Supplementary Fig. S1 belongs to one of these five patterns. (**b**) Average formations of each team throughout five games. The label "others" means that player substitutions occurred in the first half of the game, or the average formation could not be identified.

where n represents the size of C. From Eq. (5), $h(C_1, C_2)$ equals to Eq. (4) at the state where each cluster contains one Delaunay network. In addition, Ward's method yields comparable size of clusters at each hierarchy compared with other methods. (iii) The clustering process in step (ii) is displayed by the dendrogram whose vertical axis (height) corresponds to $h(C_1, C_2)$ between two merged clusters C_1 and C_2. Particular number N_c of clusters are extracted by cutting the dendrogram at a certain height h_c. (iv) Coarse-grained formations are visualized from each cluster as follows. For each Delaunay network in a cluster, the positional coordinates of each player are converted into normalized coordinates using Eq. (3). This transformation enables to compare each Delaunay network independently of $\overrightarrow{r_c}$ and $\sigma(t)$. Next, the time averaged position of each player is visualized by an ellipse in the normalized coordinates. We note that the direction and magnitude of each ellipse are determined by the eigenvector and eigenvalue of covariance matrix of each player's normalized position, $\overrightarrow{R_j}(t)$.

The use of the hierarchical clustering enables to control the number of clusters N_c according to the resolution of formations; in fact, if we want to characterize the formation changes in a short time interval, large N_c is selected and vice versa. As an example, we demonstrate the above clustering process based on Sept. 25 game of "Sendai". Figure 1(b) is the dendrogram obtained from the step (ii) where $\Delta f = 25$. The optimal number of clusters N_c can roughly be determined as the point where height increases rapidly with decreasing of the number of clusters (Fig. 1(c)); in this case, we chose $N_c = 3$. In Fig. 1(d), we show the clusters (coarse-grained formations) for $N_c = 3$ in the normalized coordinates where the direction of offense is upward. Each cluster is distinguished by a cluster number from C1 to C3, and the difference between them is that several pairs of players exchange their positions: players 2 and 3, and players 5 and 6, for example.

Results

Clustering algorithm for multiple games. Let us consider the problem of clustering Delaunay networks over multiple games. Players of a team for j-th game are identified with uniform numbers, $\overrightarrow{U_j}$. As we have shown above, the assignment of multiple uniform numbers $\overrightarrow{U_1}$, $\overrightarrow{U_2}$, … for different games to one set of indexes $\overrightarrow{I}$ of $A(t)$ is not uniquely determined. Here, we adopt the framework of "role representation" introduced by Bialkowski *et al.*[13,14]. We assume that the players play the same roles if they occupy similar positions in a formation. Then, we label each player by a role number and identify them with the indexes $\overrightarrow{I}$ of $A(t)$. In the following, we propose an extended clustering algorithm based on this idea, consisting of three parts I, II, and III.

Part I: clustering into average formation. In part I, we assign the same index i of $A(t)$ to players whose positions in a formation are approximately the same. To estimate the relative position of each player in a game, we compute the heat map of each game for each team in the normalized coordinates given by Eq. (3). We present the heat maps obtained for all teams and games in Supplementary Fig. S1. In this figure, the time-averaged position of each player is expressed by the region within each ellipse where the direction of offense is upward. The direction and magnitude of an ellipse are determined by the eigenvector and eigenvalue of the covariance matrix for the corresponding player's normalized position. These heat maps appear to be classified into several patterns. In fact, we find from our data that they belong to one of the following five patterns: "442", "4141", "433", "541", and "343" (these are referred to as "average formations" hereafter). A schematic representation of the five average formations is shown in Fig. 2(a). The frequency of such formations for each team in five games is shown in Fig. 2(b). It should be noted that we manually classified the heat maps into the average formations. Almost all teams, except the teams with player substitutions in the first half, can be classified into one of the average formations. Hence, the change in average formations during a game did not occur in our data. We also note that the names of the average formations are not an official one and other notations can also be considered.

For a certain average formation, the ellipses (average positions of players in a game) are distinguished by serial numbers from 1 to 10, as shown in Fig. 2(a). It is considered that players belonging to the same average formation with the same serial number play the same role in the team (e.g., player 1 in "4141" is interpreted as a "center

forward"). Therefore, we identify these serial numbers with the indexes $\vec{I}$ of $A(t)$, and a one-to-one correspondence between $\vec{U}_1$, $\vec{U}_2$, … and $\vec{I}$ is obtained for each average formation.

Part II: hierarchical clustering of average formations. As shown in Supplementary Fig. S1, the ellipses of some players in a heat map overlap, indicating that these players exchange their positions or move close to each other in the game. Besides, the configurations of players are slightly different even within the same average formation. In order to distinguish such patterns, in part II, we cluster all the Delaunay networks belonging to the same average formation using the clustering algorithm introduced in Methods.

Figure 3 presents typical examples of clustering results for the five games of "Sendai" where $\Delta f = 25$, with $N_c = 5$ or 15. Because "Sendai" adopted "442" in all five games (see Fig. 2(b)), the coarse-grained formation obtained using this method is expressed as "442-C1", where the former number denotes the average formation and the latter is the cluster number. Furthermore, each ellipse in a cluster in Fig. 3 consists of all the positions of players with the same index in the five games. We find that each cluster exhibits a more specific pattern compared with the corresponding average formations. The major difference between clusters is that players 2 and 3, or players 5 and 6 exchange their positions. We note that C3 in $N_c = 5$ or C7 in $N_c = 15$ include irregular patterns, which could be associated with transitional situations such as competition in front of goal or counter attacks.

The value of N_c depends on the cutting height h_c of the dendrogram, where the height represents the distance between two merged clusters in the clustering process. As noted in Methods, we can control the degree of coarse-graining of formations by varying N_c: finer (coarser) patterns are obtained by increasing (decreasing) N_c. For example, C2 in $N_c = 5$ is divided into (C2, C3, C4, C5) in $N_c = 15$; C5 in $N_c = 5$ is divided into (C11, C12, C13, C14, C15) in $N_c = 15$. In addition, when $N_c = 15$, the positions of players 7 and 10 are slightly different between clusters compared with the case of $N_c = 5$. In particular, there are two patterns that players 7 and 10 are in a middle line and a back line; such two patterns appear to correspond to the offense and defense scenes, respectively. We note that the special case, $N_c = 1$, is the most coarse pattern, corresponding to the superposition of all average formations of the five games.

Part III: transition network between clusters. When a certain number N_c of clusters is given, a continuous time series of formation changes can be regarded as discrete transitions between the clusters. In Fig. 4(a), we present transition networks, whose nodes and edges represent clusters and number of transitions between them, for the five games of "Sendai". Here, each node in the networks corresponds to the coarse-grained formation shown in Fig. 3(d), and a transition from one cluster to another represents a change of the configuration of players in the formation; e.g., C1 → C2 indicates that the players 2 and 3 exchange their positions. The nodes are placed using Fruchterman-Reingold force-directed algorithm[17], which achieves an optimal layout depending on the number of transitions between clusters (weight of edges): two nodes with a large number of transitions are placed in nearby locations. In addition, we also visualize adjacency matrices of corresponding transition networks in Fig. 4(a).

We find from Fig. 4 that each of the five games exhibits similar transition patterns as follows. First, there are two communities consisting of clusters (C1, C2, C3, C4, C5), and (C9, C10, C11, C12, C13, C14, C15); the former (latter) community corresponds to the pattern that the player 5 is on the right (left) and the player 6 is on the left (right). Second, cluster C6 is the coarse-grained formation connecting such two communities; in fact, players 5 and 6 are lined up vertically in the formation. Third, clusters C7 and C8 are somewhat irregular formations, e.g., positions of players 1 and 4 in C8 are different from other clusters. It is noted that each community includes a cluster corresponding to the position-exchanged pattern between players 2 and 3, i.e., C1 and (C9, C10). We further show the time series of the clusters for Sept. 25 game of "Sendai" in Fig. 4(b). We find that the transition between two communities occurs only a few times in the first half; namely, if players 5 and 6 exchange their positions once, the formation continues for a while. On the other hand, the duration time of the clusters C1 and (C9, C10) is not such a long, and players 2 and 3 exchange positions more frequently. Because we confirmed that such features are in common for the five games, this appears to reflect the strategy of "Sendai".

Discussion

We have proposed an extended clustering algorithm based on role representation (part I) and hierarchical clustering (part II). Here, we compare our clustering algorithm with the method introduced by Bialkowski *et al.*[13,14]. In that method, a 2D probability distribution $H(\vec{R})$ (heat map) for a team is divided into 10 heat maps, $H(\vec{R}) = \sum_{r=1}^{10} H_r(\vec{R})$, and the set $\mathcal{F} = \{H_r(\vec{R}); r = 1, \cdots, 10\}$ is regarded as the formation. Each $H_r(\vec{R})$ is computed to achieve a minimal overlap with others, under the condition that each player belongs to a different r at each frame. Because each player is labeled by a role number r instead of a uniform number u at each frame, this method is called "role representation". In the role representation approach, $H_r(\vec{R})$ consists of various players at different frames, and patterns in which two players exchange their positions are regarded as the same.

In contrast, our algorithm describes an entire heat map $H(\vec{R})$ as the sum of players' heat maps, $H(\vec{R}) = \sum_{u=1}^{10} H_u(\vec{R})$, where u denotes the uniform number. The set $\mathcal{F} = \{H_u(\vec{R}); u = 1, \cdots, 10\}$ is called a "average formation". This decomposition does not achieve the minimal overlap, namely, players with different u can exchange their positions during a game. Instead, our method distinguishes such position-exchanged patterns as different formations, based on the Delaunay method and hierarchical clustering; in particular, the quantification of a formation as the Delaunay triangulation is essential because it can incorporate the information of adjacency relationships of players. In this sense, our method realizes a more detailed characterization of formations compared with that by Bialkowski *et al.*[13,14]. Although we have only shown the results for the particular datasets, our method does not depend on the details of data.

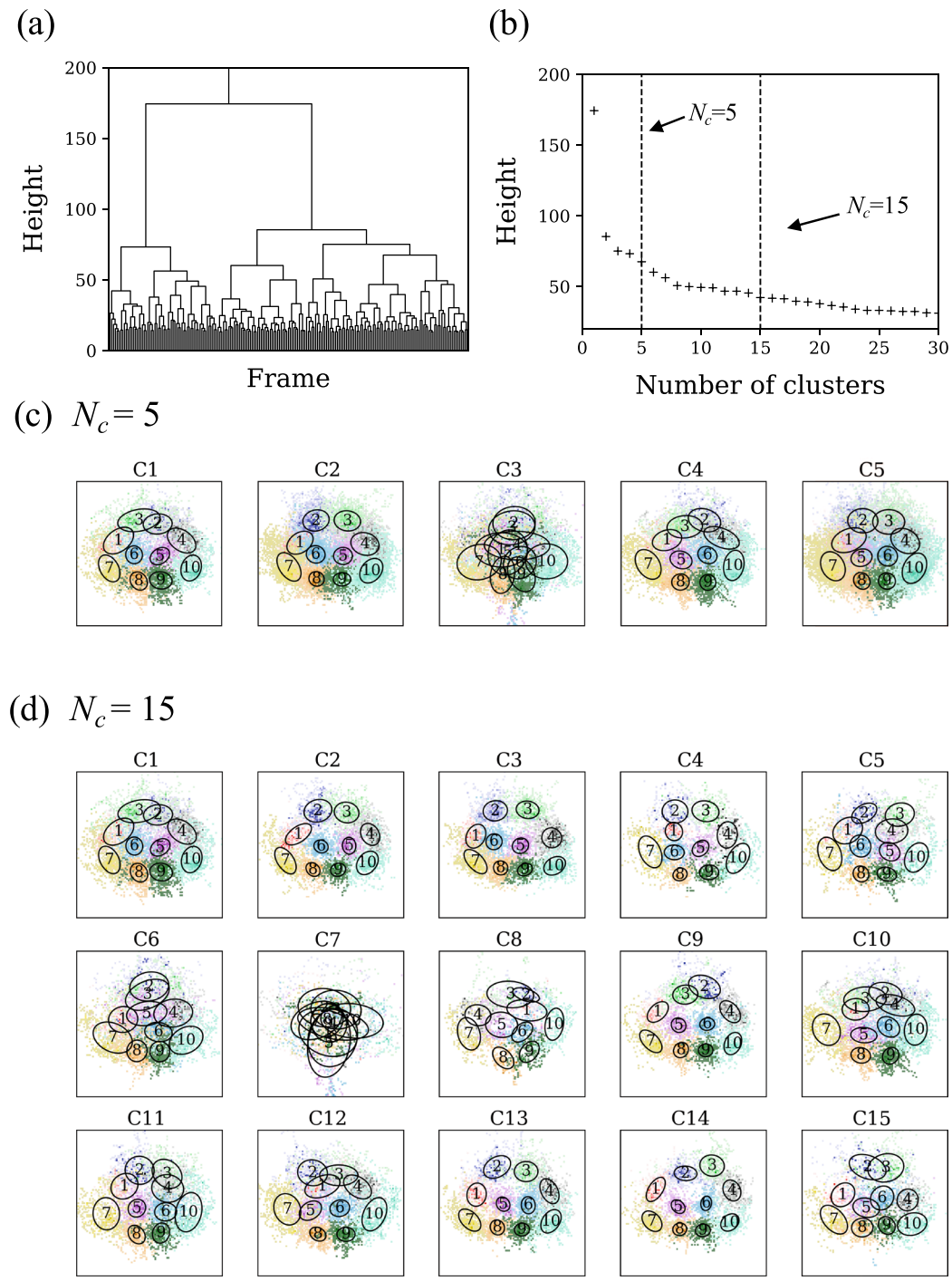

Figure 3. Results of hierarchical clustering for the five games of "Sendai". (**a**) The Dendrogram, (**b**) the relation between the number of clusters and height, and the visualization of coarse-grained formations where (**c**) $N_c = 5$, and (**d**) $N_c = 15$. Each cluster is visualized in the normalized coordinates where the direction of offense is upward and distinguished by a cluster number.

While our decomposition of the entire heat map $H(\vec{R})$ does not achieve the minimal overlap, the average positions of players, expressed by ellipses, are still clearly separated (see Supplementary Fig. S1). That is, each player carries out an individual role in a football game. This feature of football games allows us to label players not only by uniform numbers $\vec{U}$ but also by role numbers (role representation). Furthermore, it provides a criterion for the correspondence between multiple uniform numbers $\vec{U}_1$, $\vec{U}_2$, … and the indexes $\vec{I}$ of $A(t)$, and allows hierarchical clustering to be realized over multiple games. We note that our method can be applied to specific sports in which players' average positions are almost fixed because it relies on the one-to-one correspondence between $\vec{U}$ and $\vec{I}$.

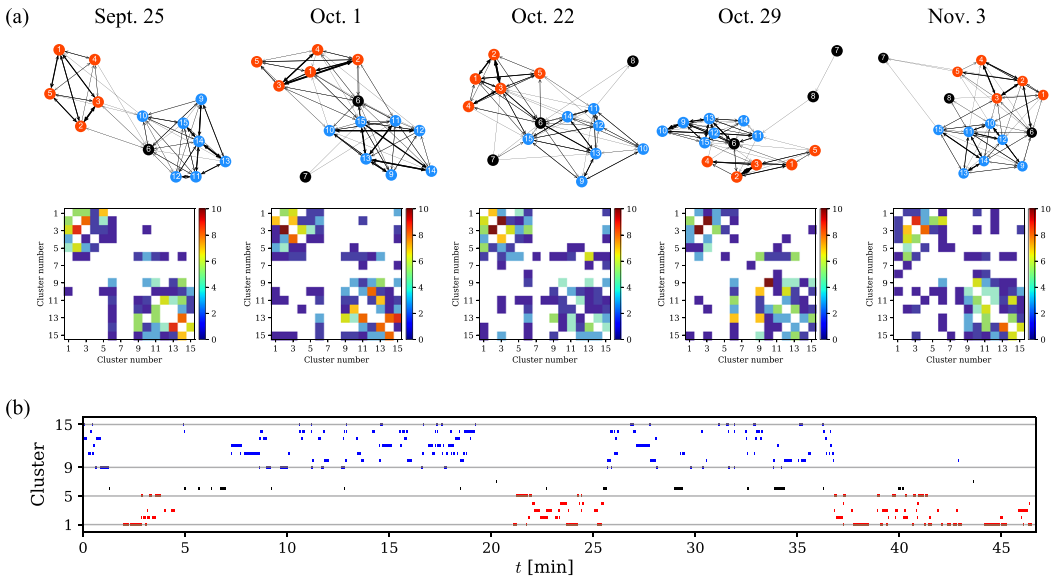

Figure 4. (**a**) Transition networks between clusters (upper panels) and those of adjacency matrices (below panels) for all games of "Sendai" where $N_c = 15$. Each node corresponding to the coarse-grained formation in Fig. 3(d) is arranged using Fruchterman-Reingold force-directed algorithm[17]. (**b**) Time series of the clusters for Sept. 25 game of "Sendai".

The variation in average formations and switches among them are a reflection of teams' strategies[1,2]. It has been reported that football teams adopt a so-called "win-stay lose-shift strategy" for formation changes between games[2]: they tend to adopt the same (a different) formation after a win (loss). Our method has the potential to provide a more detailed characterization of strategies or game flow by focusing on formation changes within a game. As an example, we have introduced the transition networks between clusters in Fig. 4. While we have mentioned some common features in Results, a closer look at the adjacency matrices in Fig. 4 shows that each network exhibits slightly different transition patterns. In order to extract more specific patterns from them, larger N_c is needed. We expect that temporal network analysis for the cluster transitions for different N_c values provide insights into the characterization of team styles.

Regarding this type of analysis, a further extension of the Delaunay network could also be considered. In fact, the present Delaunay network lacks information on opposing teams. This means that the edges do not always represent pass routes, because opposing players may exist on these edges. We can address this problem by introducing a Delaunay triangulation method including an opposing team. In this extended Delaunay network, edges connecting players in the same team represent secure pass routes. Further dynamical analyses of formation structures incorporating ball passes or interactions with opposing players by employing extended Delaunay network will be a topic of future research.

Finally, the Delaunay method and the clustering algorithm using hierarchical clustering are a general framework to coarse grain a many-particle system with incorporating its adjacency relationships. It realizes more detailed characterization and visualization rather than macroscopic quantities such as the centroid and the standard deviation for collective motions of various systems, including team sports[11], animals[18], and robots[19]. We expect that our method will provide a common tool for formation analysis of team sports and new insights to the research fields of general collective motions.

Data Availability

The dataset (player tracking data in J-League matches) was collected by DataStadium Inc., Japan, and is not publicly available because of our agreement with the company.

References

1. Hirotsu, N., Ito, M., Miyaji, C., Hamano, K. & Taguchi, A. Modeling tactical changes of formation in association football as a non-zero-sum game. *Journal of Quant. Analysis Sports* **5** (2009).
2. Tamura, K. & Masuda, N. Win-stay lose-shift strategy in formation changes in football. *EPJ Data Sci.* **4**, 9 (2015).
3. Memmert, D., Lemmink, K. A. & Sampaio, J. Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine* **47**, 1–10 (2017).
4. Sumpter, D. *Soccermatics: Mathematical adventures in the beautiful game*. (Bloomsbury Sigma, London, 2017).
5. Okabe, A., Boots, B., Sugihara, K. & Nok-Chiu, S. *Spatial tessellations: concepts and applications of Voronoi diagrams*. (John Wiley & Sons, New York, 2000).
6. Kim, S. Voronoi analysis of a soccer game. *Nonlinear Analysis: Model. Control.* **9**, 233–240 (2004).
7. Fonseca, S., Milho, J., Travassos, B. & Araújo, D. Spatial dynamics of team sports exposed by Voronoi diagrams. *Hum. Mov. Sci.* **31**, 1652–1659 (2012).
8. Taki, T., Hasegawa, J. & Fukumura, T. Development of motion analysis system for quantitative evaluation of teamwork in soccer games. *Proc. 3rd IEEE Int. Conf. on Image Process.* **3**, 815–818 (1996).

9. Taki, T. & Hasegawa, J. Visualization of dominant region in team games and its application to teamwork analysis. *Proc. Comput. Graph. Int.* **2000**, 227–235 (2000).
10. Fujimura, A. & Sugihara, K. Geometric analysis and quantitative evaluation of sport teamwork. *Syst. Comput. Jpn.* **36**, 49–58 (2005).
11. Gudmundsson, J. & Wolle, T. Football analysis using spatio-temporal tools. *Comput. Environ. Urban Syst.* **47**, 16–27 (2014).
12. Gudmundsson, J. & Horton, M. Spatio-temporal analysis of team sports. *ACM Comput. Surv. (CSUR)* **50**, 22 (2017).
13. Bialkowski, A. *et al*. Large-scale analysis of soccer matches using spatiotemporal tracking data. *Proc. 2014 IEEE Int. Conf. on Data Min*. 725–730 (2014).
14. Bialkowski, A. *et al*. Discovering team structures in soccer from spatiotemporal data. *IEEE Transactions on Knowl. Data Eng.* **28**, 2596–2605 (2016).
15. Narizuka, T. & Yamazaki, Y. (In Japanese) Characterization of the formation structure in team sports. *Proc. Inst. Stat. Math. Special Top. New Challenges to Stat. Sci. Sports* **65**, 299–307 [English version:arXiv:1802.06766] (2017).
16. Pang-Ning, T., Steinbach, M. & Kumar, V. *Introduction to data mining*. (Addison Wesley, Boston, 2005).
17. Fruchterman, T. M. & Reingold, E. M. Graph drawing by force-directed placement. *Software: Pract. experience* **21**, 1129–1164 (1991).
18. Sumpter, D. *Collective animal behavior*. (Princeton University Press, Princeton, 2010).
19. Deblais, A. *et al*. Boundaries control collective dynamics of inertial self-propelled robots. *Phys. review letters* **120**, 188002 (2018).

Acknowledgements

The authors are very grateful to DataStadium Inc., Japan for providing the player tracking data. The authors thank Hiroto Kuninaka and Tsuyoshi Mizuguchi for fruitful discussions. This work was partially supported by the Data Centric Science Research Commons Project of the Research Organization of Information and Systems, Japan, a Grant-in-Aid for Young Scientists (18K18013) from the Japan Society for the Promotion of Science (JSPS), and Hayao Nakayama Foundation for Science, Technology and Culture (H29-A2-30).

Author Contributions

T.N. designed the study and performed the analyses. Y.Y. supervised the study and proposed the direction of the analyses. T.N. prepared the manuscript, and Y.Y. checked it critically. All authors discussed the results and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-48623-1.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.