Finding Interesting Pass Patterns from Soccer Game Records

Shoji Hirano and Shusaku Tsumoto

Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
hirano@ieee.org, tsumoto@computer.org

Abstract. This paper presents a novel method for finding interesting pass patterns from soccer game records. Taking two features of the pass sequence – temporal irregularity and requirements for multiscale observation – into account, we have developed a comparison method of the sequences based on multiscale matching. The method can be used with hierarchical clustering, that brings us a new style of data mining in sports data. Experimental results on 64 game records of FIFA world cup 2002 demonstrated that the method could discover some interesting pass patterns that may be associated with successful goals.

1 Introduction

Game records of sports such as soccer and baseball provide important information that supports the inquest of each game. Good inquest based on the detailed analysis of game records makes the team possible to clearly realize their weak points to be strengthened, or, superior points to be enriched. However, the major use of the game records is limited to the induction of basic statistics, such as the shoot success ratio, batting average and stealing success ratio. Although video information may provide useful information, its analysis is still based on the manual interpretation of the scenes by experts or players.

This paper presents a new scheme of sports data mining from soccer game records. Especially, we focus on discovering the features of pass transactions, which resulted in successful goals, and representing the difference of strategies of a team by the pass strategies. Because a pass transaction is represented as a temporal sequence of the position of a ball, we used clustering of the sequences. There are two points that should be technically solved. First, the length of a sequence, number of data points constituting a sequence, and intervals between data points in a sequence are all irregular. A pass sequence is formed by concatenating contiguous pass events; since the distance of each pass, the number of players translating the contiguous passes are by nature difference, the data should be treated as irregular sampled time-series data. Second, multiscale observation and comparison of pass sequences are required. This is because a pass sequence represents both global and local strategies of a team. For example, as a global strategy, a team may frequently use side-attacks than counter-attacks.

J.-F. Boulicaut et al. (Eds.): PKDD 2004, LNAI 3202, pp. 209–218, 2004.

As a local strategy, the team may frequently use one-two pass. Both levels of strategies can be found even in one pass sequence; one can naturally recognize it from the fact that a video camera does zoom-up and zoom-out of a game scene. In order to solve these problems, we employed multiscale matching [1], [2], a pattern recognition based contour comparison method.

The rest of this paper is organized as follows. Section 2 describes the data structure and preprocessing. Section 3 describes multiscale matching. Section 4 shows experimental results on the FIFA world cup 2002 data and Section 5 concludes the results.

2 Data Structure and Preprocessing

2.1 Data Structure

We used the high-quality, value-added commercial game records of soccer games provided by Data Stadium Inc., Japan. The current states of pattern recognition technique may enable us to automatically recognize the positions of ball and players [3], [4], [5], however, we did not use automatic scene analysis techniques because it is still hard to correctly recognize each action of the players.

The data consisted of the records of all 64 games of the FIFA world cup 2002, including both heats and finals, held during May-June, 2002. For each action in a game, the following information was recorded: time, location, names(number) of the player, the type of event (pass, trap, shoot etc.), etc. All the information was generated from the real-time manual interpretation of video images by a well-trained soccer player, and manually stored in the database. Table 1 shows an example of the data. In Table 1, 'Ser' denotes the series number, where a series denotes a set of contiguous events marked manually by expert. The remaining fields respectively represent the time of event occurrence ('Time'), the type of event ('Action'), the team ID ('T_1') and player ID (P_1) of one acting player 1, the team ID ('T_2') and player ID (P_2) of another acting player 2, spatial position of player 1 (X_1, Y_1), and spatial position of player 2 (X_1, Y_1), Player 1 represents the player who mainly performed the action. As for pass action, player 1 represents the sender of a pass, and player 2 represents the receiver of the pass. Axis X corresponds to the long side of the soccer field, and axis Y corresponds

Table 1. An example of the soccer data record.

Ser	Time	Action	T_1	P_1	T_2	P_2	X_1	Y_1	X_2	Y_2
1	20:28:12	KICK OFF	Senegal	10			0	-33		
1	20:28:12	PASS	Senegal	10	Senegal	19	0	-50	-175	50
1	20:28:12	TRAP	Senegal	19			-175	50		
1	20:28:12	PASS	Senegal	19	Senegal	14	-122	117	3004	451
1	20:28:14	TRAP	Senegal	14			3004	451		
⋮				⋮						
169	22:18:42	P END	France	15			1440	-685		

to the short side. The origin is the center of the soccer field. For example, the second line in Table 2 can be interpreted as: Player no. 10 of Senegal, locating at (0,-50), sent a pass to Player 19, locating at (-175,50).

2.2 Preprocessing

We selected the series that contains important 'PASS' actions that resulted in goals as follows.

1. Select a series containing an 'IN GOAL' action.
2. Select a contiguous 'PASS' event. In order not to divide the sequence into too many subsequences, we regarded some other events as contiguous events to the PASS event; for example, TRAP, DRIBBLE, CENTERING, CLEAR, BLOCK. Intercept is represented as a PASS event in which the sender's team and receiver's team are different. However, we included an intercept into the contiguous PASS events for simplicity.
3. From the Selected contiguous PASS event, we extract the locations of Player 1, X_1 and Y_1, and make a time series of locations $p(t) = \{(X_1(t), Y_1(t))|1 \leq t \leq T\}$ by concatenating them. For simplicity, we denote $X_1(t)$ and $Y_1(t)$ by x(t) and y(t) respectively.

Figure 1 shows an example of spatial representation of a PASS sequence generated by the above process. Table 2 provides an additional information, the raw data that correspond to Figure 1. In Figure 1 the vertical line represents the axis connecting the goals of both teams. Near the upper end (+5500) is the goal of France, and near the lower end is the goal of Senegal. This example PASS sequence represents the following scene: Player no. 16 of France, locating at (-333,3877), send a pass to player 18. Senegal cuts the pass at near the center of the field, and started attack from the left side. Finally, Player no. 11 of Senegal made a CENTERING, and after several block actions of France, Player no. 19 of Senegal made a goal.

By applying the above preprocess to all the IN GOAL series, we obtained N sequences of passes $P = \{p_i|1 \leq i \leq N\}$ that correspond to N goals, where i of p_i denote the i-th goal.

3 Multiscale Comparison and Grouping of the Sequences

For every pair of PASS sequences $\{(p_i, p_j) \in P|1 \leq i < N, i < j \leq N\}$, we apply multiscale matching to compare their dissimilarity. Based on the resultant dissimilarity matrix, we perform grouping of the sequences using conventional hierarchical clustering [6].

Multiscale Matching is a method to compare two planar curves by partly changing observation scales. We here briefly explain the basic of multiscale matching. Details of matching procedure are available in [2].

Let us denote two input sequences to be compared, p_i and p_j, by A and B. First, let us consider a sequence $x(t)$ containing X_1 values of A. Multiscale

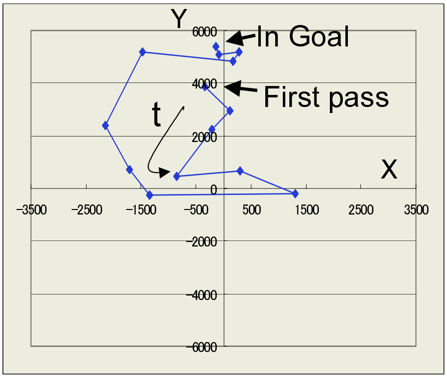


Fig. 1. Spatial representation of a PASS sequences.

representation of $x(t)$ at scale σ, $X(t,\sigma)$ can be obtained as a convolution of $x(t)$ and a Gaussian function with scale factor σ as follows.

$$X(t,\sigma) = \int_{-\infty}^{+\infty} x(u) \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-u)^2/2\sigma^2} du \tag{1}$$

where the gauss function represents the distribution of weights for adding the neighbors. It is obvious that a small σ means high weights for close neighbors, while a large σ means rather flat weights for both close and far neighbors. A sequence will become more flat as σ increases, namely, the number of inflection points decreases. Multiscale representation of $y(t)$, $Y(t,\sigma)$ is obtained similarly.

The curvature of point t at scale σ is obtained as follows.

$$K(t,\sigma) = \frac{X'Y'' - X''Y'}{(X'^2 + Y'^2)^{3/2}}, \tag{2}$$

where X', X'', Y' and Y'' denote the first- and second-order derivatives of $X(t,\sigma)$ and $Y(t,\sigma)$ by t, respectively. The m-th order derivative of $X(t,\sigma)$, $X^{(m)}(t,\sigma)$, is derived as follows.

$$X^{(m)}(t,\sigma) = \frac{\partial^m X(t,\sigma)}{\partial t^m} = x(t) \otimes g^{(m)}(t,\sigma). \tag{3}$$

Figure 2 provides an illustrative example of multiscale description of A.

Next, we divide the sequence $K(t,\sigma)$ into a set of convex/concave subsequences called segments based on the place of inflection points. A segment is a subsequence whose ends correspond to the two adjacent inflection points, and can be regarded as a unit representing substructure of a sequence.

Let us assume that a pass sequence $A^{(k)}$ at scale k is composed of R segments. Then $A^{(k)}$ is represented by

$$A^{(k)} = \left\{ a_i^{(k)} \mid i = 1, 2, \cdots, R^{(k)} \right\}, \tag{4}$$

Table 2. Raw data corresponding the sequence in Figure 1.

Ser	Time	Action	T_1	P_1	T_2	P_2	X_1	Y_1	X_2	Y_2
47	20:57:07	PASS	France	16	France	18	-333	3877	122	-2958
47	20:57:08	PASS	France	18	France	17	122	2958	-210	-2223
47	20:57:10	DRIBBLE	France	17			-210	2223	-843	-434
47	20:57:14	PASS	France	17	France	4	-843	434	298	-685
47	20:57:16	PASS	France	4	France	6	298	685	1300	217
47	20:57:17	TRAP	France	6			1300	217		
47	20:57:19	CUT	Senegal	6			-1352	-267		
47	20:57:19	TRAP	Senegal	6			-1352	-267		
47	20:57:20	PASS	Senegal	6	Senegal	11	-1704	702	-2143	2390
47	20:57:21	DRIBBLE	Senegal	11			-2143	2390	-1475	5164
47	20:57:26	CENTERING	Senegal	11			-1475	5164		
47	20:57:27	CLEAR	France	17			175	4830		
47	20:57:27	BLOCK	France	16			281	5181		
47	20:57:27	CLEAR	France	16			281	5181		
47	20:57:28	SHOT	Senegal	19			-87	5081		
47	20:57:28	IN GOAL	Senegal	19			-140	5365		

where $a_i^{(k)}$ denotes the i-th segment of $A^{(k)}$ at scale $\sigma^{(k)}$. By applying the same process to another input sequence B, we obtain the segment-based representation of B as follows.

$$B^{(h)} = \left\{ b_j^{(h)} \mid j = 1, 2, \cdots, S^{(h)} \right\} \tag{5}$$

where $\sigma^{(h)}$ denote the observation scale of B and $S^{(h)}$ denote the number of segments at scale $\sigma^{(h)}$.

The main procedure of multiscale matching is to find the best set of segment pairs that minimizes the total segment difference. Figure 3 illustrates the process. For example, five contiguous segments A_1 - A_5 at the lowest scale of Sequence A are integrated into one segment A_6 at the middle scale, and the integrated segment A_6 well matches to one segment B_1 in Sequence B at the lowest scale. Thus the set of the five segments in Sequence A and the one segment in Sequence B will be considered as a candidate for corresponding subsequences. While, another pair of segments A_0 and B_0 will be matched at the lowest scale. In this way, matching is performed throughout all scales.

There is one restriction in determining the best set of the segments. The resultant set of the matched segment pairs must not be redundant or insufficient to represent the original sequences. Namely, by concatenating all the segments in the set, the original sequence must be completely reconstructed without any partial intervals or overlaps. The matching process can be fasten by implementing dynamic programming scheme [2].

Dissimilarity $d(a_i^{(k)}, b_j^{(h)})$ of two segments $a_i^{(k)}$ and $b_i^{(h)}$ is defined as follows.

$$d(a_i^{(k)}, b_j^{(h)}) = \frac{\mid \theta_{a_i}^{(k)} - \theta_{b_j}^{(h)} \mid}{\theta_{a_i}^{(k)} + \theta_{b_j}^{(h)}} \left| \frac{l_{a_i}^{(k)}}{L_A^{(k)}} - \frac{l_{b_j}^{(h)}}{L_B^{(h)}} \right| \tag{6}$$

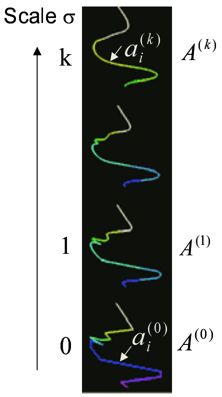


Fig. 2. An example of multiscale description.

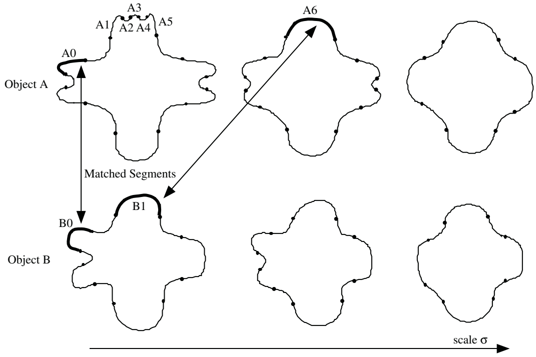


Fig. 3. An illustrative example of multiscale matching.

where $\theta_{a_i}^{(k)}$ and $\theta_{b_j}^{(h)}$ denote rotation angles of tangent vectors along segments ${a_i}^{(k)}$ and ${b_j}^{(h)}$, $l_{a_i}^{(k)}$ and $l_{b_j}^{(h)}$ denote the length of segments, $L_A^{(k)}$ and $L_B^{(h)}$ denote the total length of sequences$A$ and B at scales $\sigma^{(k)}$ and $\sigma^{(h)}$, respectively.

The total difference between sequences A and B is defied as a sum of the dissimilarities of all the matched segment pairs as

$$D(A,B) = \sum_{p=1}^{P} d(a_p^{(0)}, b_p^{(0)}), \tag{7}$$

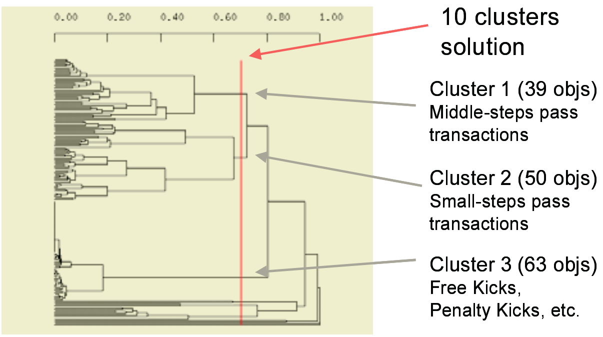
where P denotes the number of matched segment pairs.

4 Experimental Results

We applied the proposed method to the action records of 64 games in the FIFA world cup 2002 described in Section 2. First let us summarize the procedure of experiments.

1. Select all IN GOAL series from original data.
2. For each IN GOAL series, generate a time-series sequence containing contiguous PASS events. In our data, there was in total 168 IN GOAL series excluding own goals. Therefore, we had 168 time-series sequences, each of which contains the sequence of spatial positions $(x(t), y(t))$.
3. For each pair of the 168 sequences, compute dissimilarity of the sequence pair by multiscale matching. Then construct a 168 × 168 dissimilarity matrix.
4. Perform cluster analysis using the induced dissimilarity matrix and conventional agglomerative hierarchical clustering (AHC) method.

The following parameters were used in multiscale matching: the number of scales = 30, scale interval = 1.0, start scale = 1.0. In order to elude the problem of



10 clusters solution

Cluster 1 (39 objs)
Middle-steps pass transactions

Cluster 2 (50 objs)
Small-steps pass transactions

Cluster 3 (63 objs)
Free Kicks,
Penalty Kicks, etc.

Fig. 4. Dendrogram obtained by average-linkage AHC.

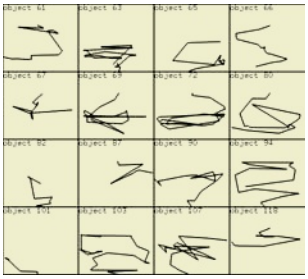


Fig. 5. Example sequences in cluster 1.

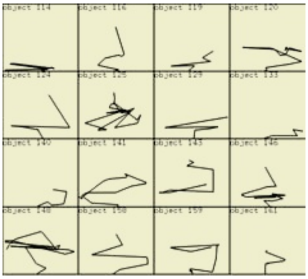


Fig. 6. Example sequences in cluster 2.

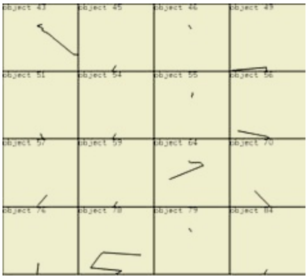


Fig. 7. Example sequences in cluster 3.

shrinkage at high scales, the shrinkage correction method proposed by Lowe et al. [7] was applied.

Figure 4 provides a dendrogram generated obtained using average-linkage AHC. Thirteen clusters solution seemed to be reasonable according to the step width of dissimilarity. However, from visual inspection, it seemed better to select 10 clusters solution, because the feature of clusters was more clearly observed. Figure 5 - 7 provide some examples of sequences clustered into the three major clusters 1, 2, 3 in Figure 4, respectively. Cluster 1 contained complex sequences, each of which contained many segments and often included loops. These sequences represented that the goals were succeeded after long, many steps of pass actions, including some changes of the ball-owner team. On the contrary, cluster 2 contained rather simple sequences, most of which contained only several segments. These sequences represented that the goals were obtained after interaction of a few players. Besides, the existence of many long line segment implied the goals might be obtained by fast break. Cluster 3 contained remarkably short sequences. They represented special events such as free kicks, penalty kicks and corner kicks, that made goals after one or a few touches. These observations demonstrated that the sequences were clustered according to the steps/complexity of the pass pass routes.

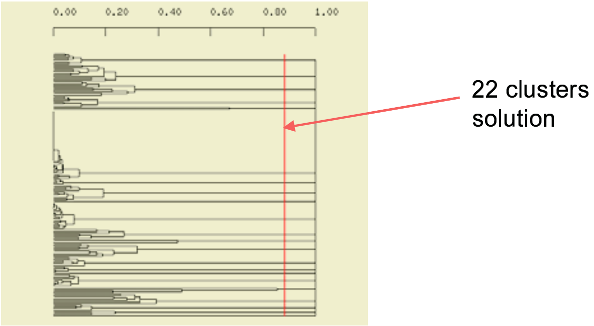


Fig. 8. Dendrogram obtained by complete-linkage AHC.

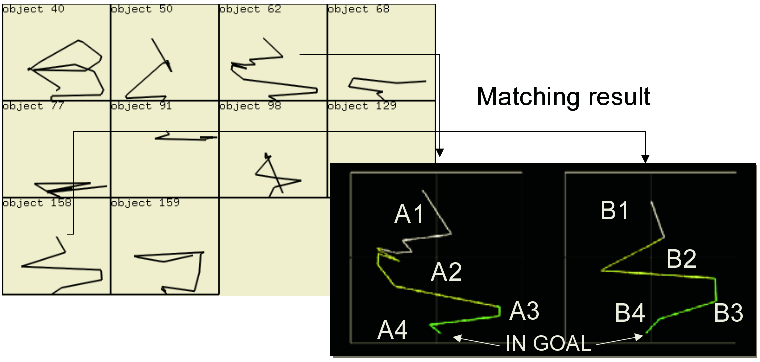


Fig. 9. Example sequences in cluster 15.

Figure 4 provides a dendrogram generated obtained using complete-linkage AHC. Because we implemented multiscale matching so that it produces a pseudo, maximum dissimilarity if the sequences are too different to find appropriate matching result, some pairs of sequences were merged at the last step of the agglomerative linkage. This gave the dendrogram in Figure 4 a little unfamiliar shape. However, complete-linkage AHC produced more distinct clusters than average-linkage AHC.

Figure 9 - 11 provide examples of sequences clustered into the major clusters 15 (10 cases), 16 (11 cases) and 19 (4 cases), for 22 clusters solution. Most of the sequences in cluster 15 contained sequences that include cross-side passes. Figure 9 right represents a matching result of two sequences in this cluster. A matched segment pair is represented in the same color, with notation of segment number A1-B1, A2-B2 etc. The result demonstrates that the similarity of pass patterns - right (A1-B1), cross (A2-B2), centering (A3-B3), shoot (A4-B4) were successfully captured. Sequences in cluster 16 contained loops. Figure 10 right shows a matching result of two sequences in this cluster. Although the directions of goals were different in these two sequences, correspondence between the loops, cross-side passes, centerings and shoots are correctly captured. This is because

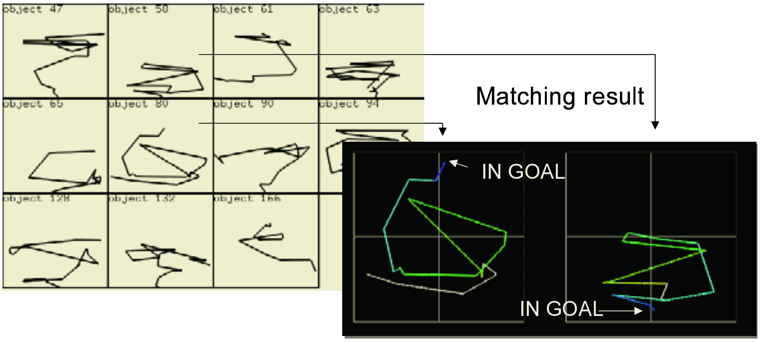


Fig. 10. Example sequences in cluster 16.

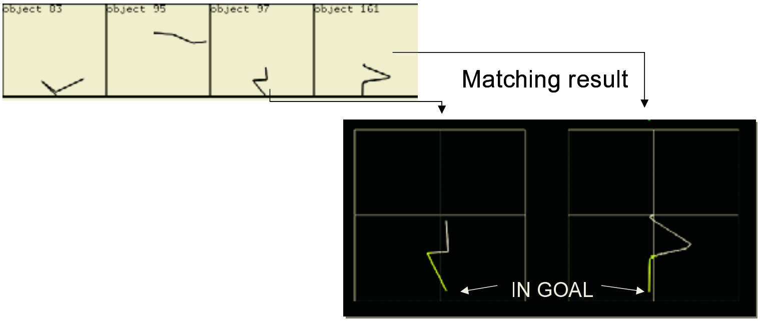


Fig. 11. Example sequences in cluster 19.

multiscale matching is invariant for affine transformations. Cluster 19 contained short step sequences. The correspondence of the segments were also successfully captured as shown in Figure 11.

5 Conclusions

In this paper, we have presented a new method for finding interesting pass patterns from time-series soccer record data. Taking two characteristics of the pass sequence – irregularity of data and requirements of multiscale observation – into account, we developed a cluster analysis method based on multiscale matching, which may build a new scheme of sports data mining. Although the experiments are in the preliminary stage and subject to further quantitative evaluation, the proposed method demonstrated its potential for finding interesting patterns in real soccer data.

Acknowledgments

The authors would like to express their sincere appreciation to Data Stadium Inc. for providing valuable dataset. This research was supported in part by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), #16700144, 2004.

References

1. F. Mokhtarian and A. K. Mackworth (1986): Scale-based Description and Recognition of planar Curves and Two Dimensional Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(1): 24-43
2. N. Ueda and S. Suzuki (1990): A Matching Algorithm of Deformed Planar Curves Using Multiscale Convex/Concave Structures. IEICE Transactions on Information and Systems, J73-D-II(7): 992–1000.
3. A. Yamada, Y. Shirai, and J. Miura (2002): Tracking Players and a Ball in Video Image Sequence and Estimating Camera Parameters for 3D Interpretation of Soccer Games. Proceedings of the 16th International Conference on Pattern Recognition (ICPR-2002), 1:303–306.
4. Y. Gong, L. T. Sin, C. H. Chuan, H. Zhang, and M. Sakauchi (1995): Automatic Parsing of TV Soccer Programs. Proceedings of the International Conference on Multimedia Computing and Systems (ICMCS'95), 167–174.
5. T. Taki and J. Hasegawa (2000): Visualization of Dominant Region in Team Games and Its Application to Teamwork Analysis. Computer Graphics International (CGI'00), 227–238.
6. B. S. Everitt, S. Landau, and M. Leese (2001): Cluster Analysis Fourth Edition. Arnold Publishers.
7. Lowe, D.G (1989): Organization of Smooth Image Curves at Multiple Scales. International Journal of Computer Vision, 3:119–130.