Proceedings of the 5th WSEAS Int. Conf. on Signal Processing, Robotics and Automation, Madrid, Spain, February 15-17, 2006 (pp307-312)

An Event Detection framework in video sequences Based on Hierarchic Event Structure Perception

LANG CONGYAN, XU DE
Institute of Computer Science, Beijing Jiaotong University, Beijing, 100044, China
E-mail: gltree@263.net, xd@computer.njtu.edu.cn

Abstract: - In this paper, we propose a framework for event detection based on hierarchic event structure perception. In order to modeling and recognizing semantic event, it is necessary to organize the spatial and temporal visual information into a meaningful representation. The main purpose of this paper is not to detect event in special domain, but to construct general event detection framework in perceptual manner and to provide meaningful unit in different semantic granularity. Specially, in the first stage fine-grained segmentation is preformed by bottom up processing that characteristics of salient regions serve as direct cues to identify temporal boundaries. Furthermore, top-down recognition module detects coarse-grain event by using HMMs to combine prior knowledge with spatio-temporal descriptors of fine-grain unit. The experimental results using different types of video sequences are presented to demonstrate the efficiency and accuracy of our proposed algorithm.

Key-Words: - Event Detection, Event Structure Perception, Salient Region, Visual Attention, Content Representation, Temporal Segmentation

1 Introduction

With the rapid growth of multimedia data, new content-based video services needs to be constructed based on semantic entities such as retrieval and manipulation of semantic content of scenes or specific event in video sequences. Toward meeting this challenge, event detection in video sequences has attracted increasing attention in recent years.

In the last several years there has been tremendous growth in the amount of computer vision research aimed at understanding human activities[1],[2]. Given a number of pre-defined actions(e.g., walking, run, etc), the problem can be stated as that of classifying a test video sequence into one of these actions. It is common in these approaches to develop a specialized parametric model for each activity, match an observed activity to all models, and choose the model that explains it best. For atomic action detection (e.g., pick up, push, sitting, etc), in [1] parameterized models of body part motions were constructed by using tracking data. Masoud et.al[2] used motion features represented by a feature image, then matching is performed in a lower dimension space using PCA.

Recently there has been a wide interest in recognition and interpretation for complex activities, such as human interaction[3], car activities in highway scenes[4]. Due to its robust against the variation of the temporal segmentation of event, Hidden Markov Models (HMMs) has been extensively applied to semantic event recognition [4],[5],[6]. In this case, a set of hidden states was specified a priori and examples were used to estimate the transition probabilities between states. For more complex activities, the approaches based on multi-agent events[3],[7] were proposed in order to model temporal pattern of multi object or motion blob.

Due to clear semantic structure in sports and news video sequences, many research efforts have been attracted on sports and news domain. Most of the approaches [8],[9] were proposed by combination of low-level features and domain knowledge. More recently, in order to narrow the semantic gap between low-level features and semantic concepts, a lot of work concentrates on the fusion of multimodal analysis. For example, in [10] multimodal information including closed caption text, speech, sound, camera motion and visual scene are integrated to detect sports event. The obvious challenge with such an approach is computation complexity of multimodal information extraction and techniques of fusion of multimodal information.

Although these systems have been successful in their respective applications, due to a wide variety of video content, it is still difficult to extract automatically semantic event from raw video data. Toward overcoming the above limitations, the main goal of this work is to construct a general event detection framework in perceptual manner. As the results, efficient semantic units can be obtained in different semantic granularity as good content pattern for high-level applications. Our approach is closely related to the work by the cognitive psychologists

Zacks[12] in the sense that perceptual-oriented event detection was made possible by the use of hierarchal event structure perception by the human visual system. The results suggested by Zacks[12] is that event can be regarded as objects in manifold of the three dimensions of space plus the one dimension of time and be perceived by a observer to have a beginning and an end. The concise interpretation for hierarchal event structure perception is that fine-grained boundaries are determined by bottom-up processing of sensory characteristics, but the grouping fine-grained segments into coarse-grained segments is modulated by top-down processing based on prior knowledge, such as intentions. Motivated by the facts, we propose a simple and general framework based on hierarchical event structure perception.

The organization of the paper is as follows: A brief overview of the proposed framework is presented in Section 2. In Section 3, we present algorithm for fine-grain unit segmentation. Based on the content unit extracted in the first stage, semantic event detection in soccer domain using HMMs is presented in section 4. The effectiveness of the proposed approach is verified by the experiments on several video sequences in the section 5. Concluding remarks are given in section 6.

2 System Overview

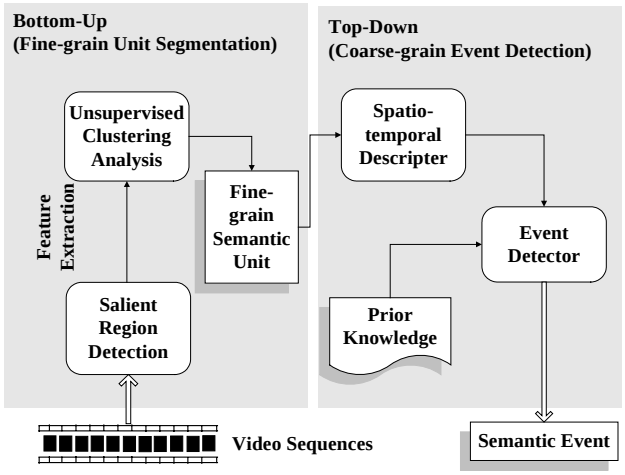


Fig.1 System Flowchart of event detection framework

Architecture of our framework is shown in figure 1. Different from putting the emphasis on the mutual interaction between different levels of representation, the research focus of this paper is to detect semantic event by bottom-up and top-down two information processing modules respectively. In particular, fine-grain unit detection is preformed by bottom up processing that sensory characteristics serve as direct cues to identify temporal boundaries. Furthermore, top-down recognition module detects coarse temporal grain, semantic event in specific domain, by combining prior knowledge with spatio-temporal descriptors of fine-grain unit.

The result of event detection is that two-level semantic content is created in a content hierarchy from the fine-grained unit to the coarse-grain unit. In this paper, the scheme for event detection is based on our prior work[13], which provided a robust salient region extraction for each frame.

3 Fine-grain unit Segmentation

This section aims at temporally fine-grained segmentation for continuous video streams. In our prior work[13], the method of salient region detection was presented, which segment spatially each frame into two parts: salient region (*SR*) and non-salient region (*NSR*). In Ref. [12], they asked human observers to independently mark the boundaries of the smallest meaningful unit of input video sequences, our expectation is that this smallest meaningful unit by human observers is likely to be correlated with fine-grain unit segmented by our methods. Thus, after the fine-grained segmentation for continuous video streams, large-scale semantic event in specific domain can be detected based on semantic features and perceptual content pattern.

3.1 Fine-grain Unit Segmentation Based On Mean Shift Clustering

In this stage, similar spatial information for each frame should be grouping into temporal homogenous meaningful unit in a bottom-up manner. The mean shift procedure[11] has been shown to have excellent properties for clustering and mode-detection with real data. The details can be consulted for the literature[11]. Thus, we exploit mean shift clustering to perform the segmentation of fine-grain unit.

Based on the salient regions extracted, feature vector can be constructed easily and efficiently for describing the content of frames. We use color and motion information to construct feature vectors. More specially, color information is calculated as the color histogram in HSV space (8-bins:H, 4-bins:S, 4-bins:V), denoted as X^c , and average motion intensity X^m for the salient regions is used, and then feature vector for each frame is denoted as

$$X^f = \{W_c \overline{X}^c, W_m \overline{X}^m\} \quad (1)$$

where $\overline{X}^c, \overline{X}^m$ denote the normalized feature vectors of color and motion , respectively. And two normailize feature vectors are fused weightly into one feauture vector X^f to represent the image

features for each frame. The weight W_c, W_m account for the reliability of each feature, hence, we set higher value of the weight W_c for color ($W_c = 0.65, W_m = 0.35$).

Since the purpose of our work in this stage is temporally segment continuous video stream into small meaningful unit, different from most existing methods of mean shift clustering, temporal range is an important cue for limitation of temporal distances between two frames. So, in our work, temporal distance is integrated with spatial information to clustering salient regions that are homogenous content and temporally adjacent. Thus, the multivariate kernel is defined as the product of three radially symmetric kernels and the Euclidean metric allows a single bandwidth parameter for each domain

$$K(X) = \frac{C}{h_f^2 h_t^2} k\left(\left\|\frac{X^f}{h_f}\right\|\right) k\left(\left\|\frac{X^t}{h_t}\right\|\right) \qquad (2)$$

where X^f is the image feature part, X^t is the temporal range, k(x) the common profile used in both two domains, h_f and h_t are employed kernel bandwidths, and the C is the corresponding normalization constant. In practice, a normal kernel can provide satisfactory performance, so the user only has to set the bandwidth parameter $h=\{h_f, h_t\}$.

A robust nonparametric clustering of the data is achieved by applying the mean shift procedure to feature space, which provides a means to analyze the feature space without making arbitrary assumptions. As a result, initial fine-grain unit can be achieved corresponding to the each cluster. Therefore, the continuous video stream is parsed in the temporal domain into short video unit, each of which contains consistent spatial visual content.

3.2 Merging process

Based on the above clustering process, initial fine grain unit is obtained as the results of the clustering. Suppose that the output of the mean shift clustering for video sequence consists of N clusters, $\{C_1, C_2, \cdots, C_N\}$. Then, continuous video sequences SV can be temporal segmented as a set of fine grain unit with a cluster label, denoted as

$$SV = \{([t_b^1, t_e^1], C_i), \cdots, ([t_b^L, t_e^L], C_j)\} \qquad (3)$$
$$C_i, C_j \in \{C_1, \cdots, C_N\}$$

where t_b^k, t_e^k is beginning frame number and ending frame number for k-th initial fine-grain unit. The sequence S consists of L fine-grain units $\{u_1, \cdots, u_L\}$, each unit u_i is represented as $u_i = ([t_b^i, t_e^i], C_k)$.

The goal of the merging process is to reduce outliers for clustering and efficiently avoids the over-segmentation for fine grain unit by eliminating small units.

Given an initial fine grain unit u_i, the length $len_i = t_e^i - t_b^i < T_{\min l}$ ($T_{\min l} = 30$), its two temporally neighboring fine grain units denoted as u_{i-1}, u_{i+1}, corresponding cluster labels are C_j, C_k, respectively. The merging process can be described as follows.

1) If $C_j = C_k$, then merging the fine-grain unit u_i with the adjacent fine-grain units into new one whose length is sum of the length of three fine-grain units.

2) If $C_j \neq C_k$, then compute the similarity between u_i and u_{i-1}, u_{i+1} respectively, merging the fine-grain unit u_i with the fine-grain unit corresponding to maximum of similarity. The measure of similarity is defined as:

$$Sim(u_i, u_j) = \sum_{p=1}^{S_i} \max_{q \in S_j} \{\frac{Dis(SR_p^i, SR_q^j)}{1 + d_{pq}^2}\} \qquad (4)$$

For measure the similarity of fine-grain unit, feature of each fine-grain unit is taken from the frame corresponding to the center of cluster, called center frame. And the number of salient regions within the center frame of the fine-grain unit u_i is S_i, that is, the center frame of the fine-grain unit u_i consists of a set of salient regions $\{SR_1^i, \cdots SR_{S_i}^i\}$. Then, in the formula (4), $Dis(SR_p^i, SR_q^j)$ is Euclidian distance between feature vectors of two salient regions, and d_{ij} represents the distance between the centroids of the salient regions.

The experiment results illustrated in the section 5 verified the efficiency and reliableness of the algorithm of fine-grained segmentation. In this stage, continuous video stream is parsed in the temporal domain into short content units, each of which contains consistent visual content. It is worth pointing out that fine-grain unit extraction is based on bottom-up processing, therefore, the fine-grain unit is a general content pattern to represent video data in a meaningful form and can be flexibly adapted for more high-level content description, such as video object extraction.

4 Event Detection using HMMs

In this section, based on fine-grain unit segmented in the above bottom-up processing, we developed a method to identify large-scale event in the special domain video: sports video. The event of interest is goal event in soccer videos, which is important information for soccer video analysis. Due to the importance of temporal context information, hidden Markov models (HMMs) are a natural and simple way to model temporal relationships. In this way, we use a HMM model for our purpose. The model topology is derived from the observations, reflecting the nature of the target patterns.

Based on fine-grained segmentation, video sequence is decomposed into homogenous semantic unit. Hence, large-scale event detection is conducted directly on grouping the fine-grain units, which avoid difficulty in detecting boundaries of event. Since salient regions in the each frame capture important information in a perceptual manner, color and motion information can be extracted as robust and efficient characteristics of video content.

More specifically, motion is the most important feature for capturing the semantic contents in video, especially for sports videos, we compute the 8-bins motion intensity histogram by statistic all the salient regions for each fine-grain unit. And dominant colour for all the salient regions is computed as the color information of the unit. Set colour threshold T_{DC} =0.2, dominant colour denoted as Dc_i can be constructed according to the color value that is larger than T_{DC}.

$$Dc_i = \{(C_j, H_j),\ j = 1, 2, ..., N_C\} \qquad (5)$$

where C_j is color value, and H_j is value of histogram corresponding color value C_j, and N_C is the number of dominant color. On the other hand, the type of camera view is an important syntax prior for goal event in soccer videos. Hence in our work, each fine-grain unit is annotated with semantic categories, four types of camera view {*long views, medium views, close-up views and out-of fields*}.

Based on visual features and semantic prior, Hidden Markov Models (HMMs) is used to model features extracted to detect goal event in soccer videos. The most likely sequence of states $\hat{S}$ should be computed within a model given the observation sequence $O = \{o_1, \cdots, o_n\}$, which is obtained by $\hat{S} = \arg\max_s P(S \mid O)$. The posterior state sequence probability $P(S \mid O)$ is given by

$$P(S \mid O) = \frac{P_{s_1} p_{S_1}(O_1) \prod_{t=2}^{T} p_{S_t}(O_t)\, p_{S_t|S_{t-1}}}{P(O)} \qquad (6)$$

where $S = \{q_1, \cdots, q_k\}$ is the set of discrete states, $s_t \in S$ corresponds to the state at time t. $p_{S_t|S_{t-1}}$ is the state-to-state transition probability. And the prior probabilities for initial state are P_{S_1}, the output probabilities for each state is $p_{S_t}(O_t)$.

The prior probabilities for each state and the transition matrix between states are estimated using the labeled data. In the training process, we trained (k=6) HMMs with left-to-right topologies for goal events. The initial estimations are given by training data manually labeled, and an Expectation-Maximization algorithm (EM) is then to give the final optimized estimation. Once HMM models are learned, new video sequences can be labeled into goal event by using learned HMM models.

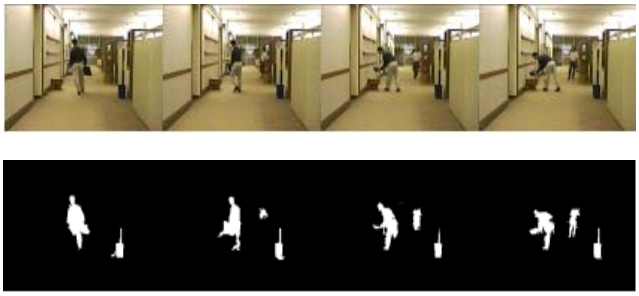
5 Experiment Results

We have tested the proposed two-stage event detection system successfully on real video streams containing a variety of video sequences: surveillance video sequences (MPEG-4 test sequences) and sports games (basketball and soccer). Due to its distinct boundaries of small-scales content change, surveillance sequences can be well used to test our first stage temporal segmentation for fine-grain unit. The event detection experiments in coarse-grain semantic granularity were conducted on the sports games. In the following we present results on (1) the results for fine-grain unit segmentation. (2)the results for even detection in soccer domain.

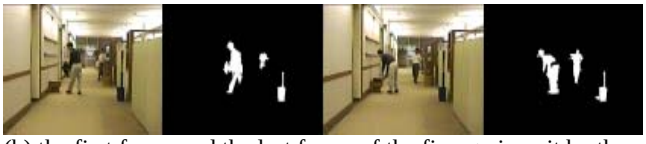
5.1 Evaluation on Fine-grain Unit Extracted

Fig.2 shows fine-grain unit segmentation results for hall-monitor video clip, where the result is selected as four clusters and represented using the center frame and corresponding salient region map of each fine-grain unit. As can be seen from the Fig.2, the salient regions have been grouped in meaningful units successfully by the process of unsupervised clustering and merging. Though the background in hall monitor sequences is stationary, there has a high level of noise present and effect by illumination. In fig.2(a), four initial fine-grain unit is formed by clustering. Due to the change of subject motion information, the last unit becomes small clusters as outliers. Then merging processing is conducted on the latter two clusters. As the result of merging, the final fine-grain unit is achieved by merging similar two clusters, fig.2(b) shows this final fine-grain unit extracted results, which is represented by its first and

last frame: raw frames and corresponding salient region map results. As expected, the boundaries of fine-grain unit segmented can reflect discontinuity of content change within video sequences.



(a) the center frames in four clusters and corresponding salient region map respectively



(b) the first frame and the last frame of the fine-grain unit by the merging the latter two clusters in (a).

Fig.2 fine-grain unit segmentation result in hall-monitor video sequences

For the psychological studies, Zacks et.al. [12] asked human observers to independently mark the boundaries of the smallest meaningful unit of input video sequences, our expectation is that this f smallest meaningful unit by human observers is likely to be correlated with fine-grain unit detected by our methods. In our experiment, we ask 10 observers to mark the boundaries of fine-grained meaningful unit, then a histogram of boundary detections is constructed as a function of the frame index. The peaks of these histograms correspond to ground truth of boundaries.

Since the main purpose of our work is to construct a general event detection framework for organizing the spatio-temporal visual information into a meaningful representation, so detected content unit boundaries in two temporal scales should exactly correspond to the real event boundaries. Hence, based on the ground truth created, similar to the criteria defined in [14], we use the measure of correlation *C* between the psychological ground truth and our results as objective evaluation criteria. The correlation *C* is calculated by the ratio between the number of boundaries correctly detected and the total numbers of boundaries marked by human observers.

Table 1 shows the part of experimental results for five test clips: hall monitor(HM), two soccer video clips(S-I, S-II) and two basketball clips(B-I, B-II). The average value 81.9% of the correlation *C* can be reached in our experiments. Specially, the boundaries of detected unit can primely correspond to the small meaningful unit marker by the users, which provides an automatic and general method for region-level spatio-temporal content unit extraction.

5.2 Event Detection Results

In our experiments, the soccer video data is MPEG-1 format at 25 per second. The model parameters of prior and conditional probabilities are obtained by statistics over a data set of 20 video clips (total time 01:14:26). The ground truth for goal event is labeled manually. The results are given for five soccer sequences, denoted as SG_I, SG_II, SG_III, SG_IV and SG_V.

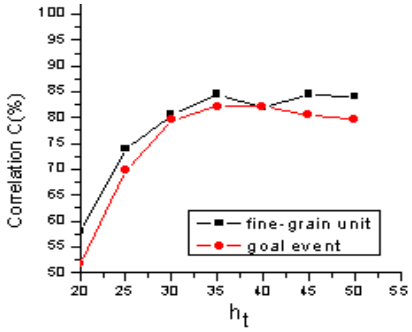


Fig.3 The correlation C-parameter h_t curve for fine-grain unit and goal event detection

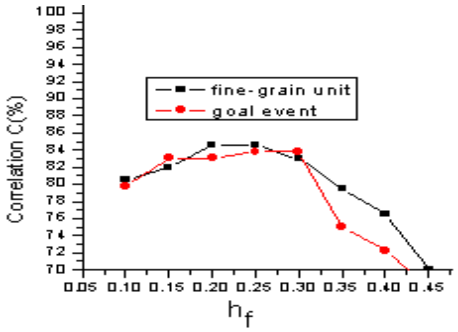


Fig.4 The correlation C-parameter h_f curve for fine-grain unit and goal event detection

In the clustering process, two bandwidth parameters h_f, h_t in spatial and temporal domain determine the scales of temporal segmentation for fine-grain unit. By the observations from the experiments, more reasonable results of our method could be achieved by setting $h_f, h_t = [0.25, 35]$. In order to evaluate the sensitivity of our event detection accuracy to variations in the two parameters, we vary h_f, h_t values to produce clusters on coarser or finer scales. Figure 3 shows the correlation *C* curve with the use of the different parameter h_t by selecting $h_f = 0.25$. And Figure 4 is correlation *C* curve with the varied parameter h_f by selecting $h_t = 35$.

The results of the experiments are summarised in Table 2. The average correlation *C* is 83.3%. Total or part occlusion, strong illumination is the main limitation of our system, the spatial information taken from salient regions cannot be matched correctly. However, the boundaries of ground truth can still be captured, which is important to robust and general event detection. The proposed method shows an encouraging improvement in semantic event detection.

6 Conclusion

In this paper, we present a semantic event detection framework based on event structure perception. The main differences of this task form traditional event detection lies in: (1) a general-purpose event detection framework is constructed to provide good spatio-temporal content pattern for variety video applications, rather than the goals of accuracy improvement for specific domain event detection. (2) Based on hierarchical event structure perception, two-level semantic granularity is extracted by bottom-up and top-down information processing respectively, which is different from putting the emphasis on the mutual interaction between different levels of representation. Future work will be focused on the learning rules generation for complicated semantic relations, which is important to high-level semantic content analysis of video sequences.

Acknowledgment

This work was supported by the Science Foundation of Beijing JiaoTong University (Grant No. 2004SM013).

References:

[1] Yaser Yacoob, *Parameterized Modeling and Recognition of Activities*, Computer Vision and Image Understanding, 73(2) (1999) 232-247

[2] Osama Masoud, Nikos Papanikolopoulos, *A method for human action recognition*, Image and Vision Computing, 21 (2003) 729-743

[3] Somboon Hongeng, Ra, Nevatia, Francois Bremond, *Video-based event recognition: activity representation and probabilistic recognition methods*, J. Computer Vision Image Understand, 96 (2004) 129-162.

[4] Medioni G., Cohen I., et al. *Event Detection and Analysis from Video Streams*. IEEE Trans. Pattern Anal. Mach. Intell, Vol.23(8), pp.873-889,2001

[5] N.Oliver, B.Rosario, A. Pentland, *A Bayesian cooputer vidson system for modeling human interactions*, IEEE Trans. Pattern Anal. Mach. Intell. 22(8),pp.831-843,2000.

[6] E.Kijak, G.Gravier, P.gros, L.Oisel and F.Bimbot, *HMM Based Strcturing of Tennis Videos Using Visual and Audio Cues*. Proc. of ICME, pp.309-312, 2003.

[7] S.Intille, A,Bobick, *Recognition planned multiperson action*, J. Computer Vision Image Understand , Vol.3,pp. 414-445, 2001.

[8] Cees G.M. Snoek, and Marcel Worring, *Multimedia Event based Video Indexing Using Time Intervals*. IEEE Trans. On Multimedia, Vol.7(4),pp.638-647, 2005.

[9] Ekin, A, Tekalp, A. M., *Automatic Soccer Video Analysis and Summarization*. IEEE Trans. On Image Processing, Vol.12(7), pp. 796-807 , 2003.

[10] M.R.Nephade, T.S.Huang, *Detecting semantic concepts using context and audio/visual features*, Proc. IEEE workshop on Detection and Recognition of Events in Video, pp.92-98,2001.

[11] D.Comaniciu, P.Meer: Mean Shift: *A Robust Approach toward Feature Space Analysis*. IEEE Trans. Pattern Anal. Mach. Intell, Vol.24(5), pp. 603-619 , 2002.

[12] J.Zacks, B.Tversky, and G.Iyer, *Perceiving, remembering, and communicating structure in Events*, Journal of Experimental Psychology: General 130(1), pp.29-58, 2001.

[13] Lang Congyan, Xu De, *Perception-Oriented Prominent Region Detection in Video Sequences*, Journal of Computing and Informatics, Vol.29(3) , pp. 253-260, 2005.

[14] Y.Rui and P.Anadan, *Segmentation visual actions based on spatio-temporal motion patterns*, Proc of IEEE Inter. Conf. On Pattern Recognition, pp.111-118, 2000.

Table 1 Statistical Results of Fine-grain Unit Detection

Video Sequences	HM	S-I	S_II	B_I	B-II
Time	00:00:09	00:03:55	00:03:02	00:02:54	00:03:11
No. initial fine-grain unit	11	29	28	33	36
No. final fine-grain unit	5	17	19	21	21
Correlation *C* (%)	75	86.6	87.3	79.6	80.9

Table 2 Statistical Results of Goal Event Detection

Video Sequences	SG_I	SG_II	SG_III	SG_IV	SG_V
No. goal events	2	2	3	2	3
No. Detected goal events	2	2	3	3	3
No. Correctly Detected boundaries	3	3	5	4	5
Correlation *C* (%)	75	75	83.3	100	83.3