Localization and Recognition of the Scoreboard in Sports Video Based on SIFT Point Matching

Jinlin Guo[1,2,4], Cathal Gurrin[1,4], Songyang Lao[2], Colum Foley[1,4], Alan F. Smeaton[1,3,4]

[1]Center for Digital Video Processing, Dublin City University, Ireland
[2]School of Information System & Management, National University of Defense Technology, China
[3]CLARITY: Centre for Sensor Web Technologies, Dublin City University, Ireland
[4]School of Computing, Dublin City University, Ireland
{jguo,cgurrin,cfoley,asmeaton}@computing.dcu.ie
laosongyang@vip.sina.com

Abstract. In broadcast sports video, the scoreboard is attached at a fixed location in the video and generally the scoreboard always exists in all video frames in order to help viewers to understand the match's progression quickly. Based on these observations, we present a new localization and recognition method for scoreboard text in sport videos in this paper. The method first matches the Scale Invariant Feature Transform (SIFT) points using a modified matching technique between two frames extracted from a video clip and then localizes the scoreboard by computing a robust estimate of the matched point cloud in a two-stage non-scoreboard filter process based on some domain rules. Next some enhancement operations are performed on the localized scoreboard, and a Multi-frame Voting Decision is used. Both aim to increasing the OCR rate. Experimental results demonstrate the effectiveness and efficiency of our proposed method.

Keywords: Localization and Recognition of Scoreboard, SIFT Points Matching, Sports Video.

1 Introduction

With the development of high-speed broadband networks and digital video technology (including generating, compression, storage and processing), the amount of sports videos to which viewers can access is increasing drastically. It's often not possible for even the most avid sports fan to watch more than a small fraction of the available coverage of a complete event, such as the World Cup. Furthermore, for many sports much of the time during an game is often not significant to the progression of the game or its outcome. Therefore automatic sports video indexing and retrieval techniques have attracted a lot of research interest. Due to the automatic indexing of sports content, users can retrieve their preferred clips of sports video such as goals in soccer.

In broadcast sports videos, a superimposed scoreboard is used to display game status such as team names, score, etc., to increase the audiences' understanding of the game progression. Furthermore, the scoreboard changes after a goal event occurs. Therefore, localization and recognition of the scoreboard is very meaningful for sports video analysis and processing, for example, as a method for detecting score events or as a source of evidence for a score detection or event detection technique.

In this paper, we present an effective and efficient method to localize and recognize the scoreboards in the videos based on the observations that the location of scoreboard is static and it is present on-screen for all the duration of the game. Firstly, a bag of matched points obtained by a modified SIFT match technique is used to represent the candidate scoreboards. Then the exact area of scoreboard is localized by computing a robust estimate of the matched points cloud in a two-stage non-scoreboard filtering process. In the recognition step, some text enhancement operations and a Multi-frame Voting Decision are performed before using a commercial OCR for increasing the OCR rate. Experiment results demonstrate the effectiveness and efficiency of our proposed method.

The rest of the paper is organized as follows. In section 2, we provide an overview of the state-of-the-art of localization and recognition of superimposed text in videos, section 3 describes the localization and recognition of scoreboard in video in detail. Section 4 presents the experiment results. Finally conclusions are drawn in section 5 and we provide an outlook for further research.

2 Related Work

Localization and recognition of superimposed text in video is a major task in video content analysis and processing. A number of algorithms to localize and recognize superimposed text from still images and video sequences have been published in recent years [1] ... [10], which can be categorized into two types: one type is localizing texts in individual image [1] ... [4], the other type is utilizing the temporality of video sequences [5] ... [10].

Jain A.K. et al. [1] employed color reduction by bit dropping and color clustering quantization firstly, and afterwards a multi-value image decomposition algorithm was applied to decompose the input image into multiple foreground and background images. Then connected component analysis was performed on each of them to localize text candidates.

Ngo C-W. et al. [2] presented a background complexities-based text detection and segmentation method, in which video frames were classified into four types according to the edge densities. Edges of the non-text regions were gradually removed by repeated shifting and smoothing operators.

In [3] and [4], the authors treated text detection as a classification problem. Xi. Li et al. [3] used SVM to obtain a text region based on the features extracted by stroke filter calculation on stroke maps. Chen D. T. et al. [4] compared the SVM-based method with multilayer perceptrons (MLP) based on text verification over four independent features, namely, the distance map feature, the gray-scale

spatial derivative feature, the constant gradient variance feature and the DCT coefficient feature. Finally they found that better detection results were obtained by using SVM rather than MLP.

In [5] Lienhart R. et al. adopted the redundant information of video frames to refine the coarse text regions detected by a pre-trained feed-forward network.

Wang R.R. et al. [6] employed a multi-frame integration method i.e. time-based minimum (or maximum) pixel value search to obtain the integrated images for the purpose of minimizing (or maximizing) the variation of the background of the image.

Tang et al. [7] proposed a universal caption detection and recognition method based on a fuzzy-clustering neural network technique.

These general methods are however either too complicated, hence time-consuming, or sensitive to selection of thresholds, and not suitable for scoreboards localization in sports video frames. Recently, texts localization and recognition in sports video has attracted some research interest.

In [8] Zhang D. et al. proposed general and domain-specific techniques. They first presented a general algorithm to detect and locate captions, and then they employed a domain model of specific sports, e.g. baseball and basketball, in the text recognition to improve its rate from 78% to 87%.

Yih-Ming et al. [9]detected and localized the text region using an iteratively temporal averaging technique in a series of sports video frames at first, and then a accurate extraction of text content was performed based on text identification and model-based segmentation processes. Finally they recognized the characters using a commercial OCR technique.

Hsieh C.H. et al. [10] proposed a detection and recognition method of scoreboard for baseball video. They firstly identified the scoreboard type using template matching and then extracted the caption region of each type. At last, the digits in the scoreboard were recognized by a neural network classifier.

A Scoreboard can be localized and recognized effectively and efficiently according to its characteristic in sports video frames. That is, the scoreboards is fixed or only slightly changed during the course of the game, namely, the font type and relative location of each field are kept the same over the whole video. Based on this observation,we present an effective and efficient method to localize and recognize scoreboards in sports videos.

3 Proposed Scoreboard Localization and Recognition Method

The method proposed in this paper consists of two processes: localization process and recognition process, as shown in Figure 1. In the localization process, it first matches SIFT points in two frames extracted from the input sports video clip. Then the scoreboard is localized based on robust estimation within a two-stage filter of non-scoreboard matched points. In the recognition process, it identifies the scoreboard text after some text enhancement operations are performed on the scoreboard. The details are described in the following sections.

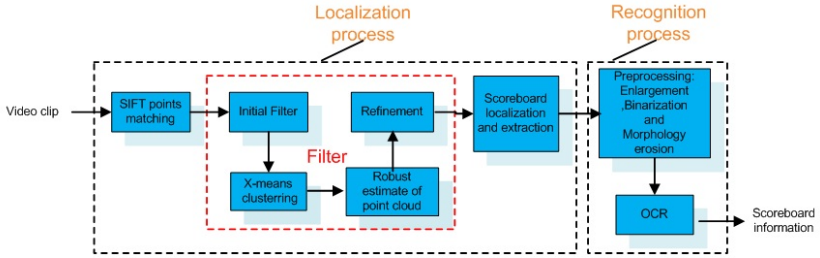

Fig. 1. Flowchart of the proposed approach

3.1 Localization Process

SIFT points Detection and Matching: Recently, it has been shown that region-based approaches are effective methods for object detection and recognition due to the fact that they can cope with the problem of occlusion and geometrical transformation [12]. These approaches are commonly based on the idea of modeling an object by a collection of local salient points. Each of these local features describe a small region around the interesting point and therefore they are robust against occlusion and clutter.

In particular, the 128-dimensional SIFT feature proposed in [11] has been proven effective in detecting objects. Because it is designed to be invariant to relatively small spatial shifts of region positions, which often occur in real images. Therefore, we use the SIFT feature as descriptor of local salient points. By combining the results of local point-based matching we are able to match an entire scoreboard.

The input video clip can be denoted as: $Clip = \{f_1, f_2, ..., f_N\}$. Here f_i denotes frame, N denotes the number of frames in this video clip.

Two frames: f_p and f_q are extracted from the input clip. It should be noted that these two frames are chosen arbitrarily for the demonstration of this method. No claim is made for any optimal frame-selection. However, these two frames should be extracted from different shots. Because of temporal redundancy, two frames from the same shot will lead to too many matched points.

SIFT points are detected on f_p and f_q using the four steps in [11], denoted as respectively:

$$\begin{aligned} T_p &= \{(x_k^p, y_k^p, s_k^p, d_k^p, o_k^p)\} \qquad for\ k \in \{1, 2, 3, \cdots, N_p\} \\ T_q &= \{(x_k^q, y_k^q, s_k^q, d_k^q, o_k^q)\} \qquad for\ k \in \{1, 2, 3, \cdots, N_q\} \end{aligned}$$

Here x_k^c, y_k^c, s_k^c, d_k^c $(c \in \{p, q\})$ are the x-position, y-position, the scale, and the dominant direction of the k_{th} SIFT point respectively. o_k^c is the 128-dimensional feature vector for each SIFT point.

So every extracted frame is represented as a bag of SIFT points. The next step is to find these matched points between two frames, i.e. points matching. The performance of points matching effects the localization of the scoreboard greatly.

Firstly we review the matching technique in [11]. Denoting P and Q as set of SIFT points for two images respectively, for any point in P, p_i ,to which q_j and $q_{j'}$ the closest and second closest Euclidean distances from points in Q. The corresponding distances are d_{ij} and $d_{ij'}$ respectively, and $d_{ij} \leq d_{ij'}$. If $d_{ij} \leq d_{ij'} * \alpha$, then p_i and q_j are matched points. α is a predefined threshold, representing the point's discrimination, in [11] the authors set $\alpha = 0.8$. According to this rule, the initial point matching between two feature point sets, in which processing, some mismatches exist, so algorithms such as RANSAC can be used to eliminate mismatches.

For the similarity measure S, if $S(p_i, q_j) = \underset{q_l \in Q}{min} S(p_i, q_l)$, then q_j is the closest point in Q,to p_i. However, if $S(p_i, q_j) \neq \underset{p_t \in P}{min} S(p_t, q_j)$, then p_i is not the closest point in P to q_j, so it's not reasonable to set p_i and q_j as matched points. The robust points matching techniques should have the feature as follows: if p_i and q_j are matched points, then $S(p_i, q_j) = \underset{q_l \in Q}{min} S(p_i, q_l)$ and $S(p_i, q_j) = \underset{p_t \in P}{min} S(p_t, q_j)$, vice-versa. Obviously, for the method in [11], $d_{ij} \leq d_{ij'} * \alpha$, $S(p_i, q_j) = \underset{q_l \in Q}{min} S(p_i, q_l)$, but not always $S(p_i, q_j) = \underset{p_t \in P}{min} S(p_t, q_j)$, so (p_i, q_j) may be mismatched points.

Based on the aforementioned analysis, we set p_i and q_j as matched points, if they satisfy as follows:

$$
\begin{aligned}
d(p_i, q_j) &= \underset{q_l \in Q}{min} (p_i, q_l) = \underset{p_t \in P}{min} (p_t, q_j) \\
d(p_i, q_j) &\leq \underset{q_l \in Q, l \neq j}{min} d(p_i, q_l) * \alpha \\
d(p_i, q_j) &\leq \underset{p_t \in P, t \neq i}{min} d(p_t, q_j) * \alpha
\end{aligned}
$$

Here $d(p_i, q_j)$ is the corresponding Euclidean Distance between p_i and q_j, and α is set as *0.8* experimentally.

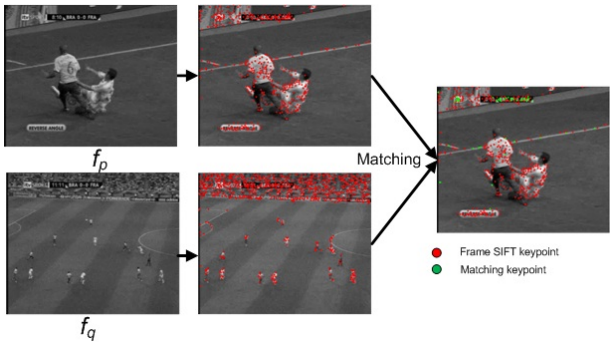

Fig. 2. SIFT points detection and matching

Based on the modified matching technique aforementioned, matched points between frame f_p and frame f_q are obtained (as shown in Figure 2). The next step is filtering some non-scoreboard matched points according to some domain-specific rules.

Initial Filtering: Certain characteristics exist when a scoreboard is shown on a video frame, which can be used to remove some non-scoreboard matched points.

- For the convenience of viewers' watching, the scoreboard always appears in the lower or upper areas of a video frame. We assume that the scoreboard always appears in the upper $1/4$ area and lower $1/4$ area. Therefore those matched points not appearing in these two areas are discarded.
- Each distance between matched points and any boundary (*top, bottom, left and right*) of the frame should be greater than T, which is a threshold and set as *15* pixels in our experiments based on observation.

As shown in Figure 3, the scoreboard always appears in either the *R1* or *R2* area. After this filtering, most of non-scoreboard matched points are removed (as shown in Figure 4(a)).

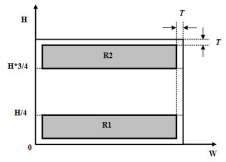

Fig. 3. Area where scoreboard is shown

Clustering and Robust Estimate of The Matched Point Cloud: After the first filtering, some non-scoreboard matched points still exist, which is caused by constant appearance of TV Logo or other objects. However, all these matched points can be clustered into one or several clusters in term of proximity of matched points generated by the same object. Clustering in this two dimensional space is performed using X-means proposed in [13]. Unlike K-means, the X-means clustering does not require the number of clusters to be predefined.

Robust Estimate is performed on each of these clusters, after which several robust centroids are localized. In this way the exact area for each cluster is obtained.

In order to localize the centroid for each cluster in the frame f_p and approximate its area, we compute a robust estimate on each matched points cluster. One matched points cluster is so denoted as $P = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$. The robust centroid estimate is computed by iteratively solving for (μ_x, μ_y) in

$$\sum_{i=1}^{n} \Psi(x_i; \mu_x) = 0, \sum_{i=1}^{n} \Psi(y_i; \mu_y) = 0$$

Here the influence function used is the Turkey biweight and the scale parameter c is estimated using the *Median Absolute Deviation (MAD)* from the median:

$$MAD_x = median_i(|x_i - median_j\,(x_j)|).$$

Refinement: After Robust Estimate, the area (represented by a rectangle, as shown in Figure 4(b)) for each cluster is localized. These rectangles whose width values are smaller than T pixels are considered as non-scoreboard and removed. In our experiment, $T = 20$ pixels based on observation. After this filtering, the scoreboard bounding box is obtained (as shown in Figure 4(c)).

Furthermore, because the scoreboard is attached at a fixed location in every frame, the localization of a scoreboard is only performed once for a video of an entire match.

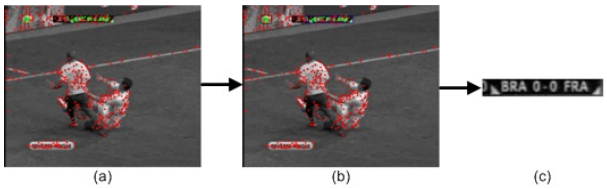

Fig. 4. SIFT matched points filtering

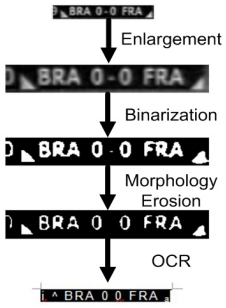

Fig. 5. Preprocessing and OCR of scoreboard text

3.2 Scoreboard Text Recognition

Current optical character recognition (OCR) techniques such as ABBYY OCR [15] or ReadIRIS [17] perform rather well and give good accuracy for texts printed on a clear background, and can recognize multiple languages by adding source character libraries. However, since we are interested in recognition of the text printed against shaded and textured backgrounds. OCR technology cannot easily handle such text. Hence we need to preprocess the extracted scoreboard before OCR so that the scoreboard can be recognized correctly and easily.

The image for the scoreboard cropped out in the localization processing is relatively simple in nature. It only contains team, score and other text, and uniform color for background (as shown in Figure 4(c)), of which team and score information is the most important. Some operations are performed on the scoreboard image before using OCR software to recognize the texts in scoreboard. Details are provided in the following section.

Preprocessing:

Step 1: Size Enlargement, to double the size of the scoreboard image by using Bicubic Interpolation [16].

The characters in scoreboard are small and compact,which need to be enlarged for increasing OCR rate. We choose bicubic interpolation due to the fact that the interpolated surface produced by bicubic interpolation is smoother than corresponding surfaces obtained by bilinear interpolation. In addition the nearest-neighbor interpolation and has fewer interpolation artifacts.

Step 2: Binarization using a threshold T obtained by the Otsu method [14].

The area of localized scoreboard mainly contains two classes of pixels background and text. Therefore, binarizing the scoreboard using the threshold T obtained by Otsu method is viable for OCR.

Step 3: Morphology Erosion [16].

A morphology erosion operation can effectively remove the noises and decrease the blur of text edges.

OCR: The commercial OCR software, designed for all alphabets, digit and symbols, of ABBYY OCR is used in our experiments. It is applied to recognize all the texts in the scoreboards (as shown in Figure 5).

Multi-frame Voting Decision: After text recognition, one result from a single frame is obtained. Because the data of each field may change after occurrence of a new score event, the text of the same field generally stays the same for a relative long time (at least 5 seconds). This characteristic can be employed to further improve the recognition rate. In this work, we use the majority voting technique for several consecutive frames to correct the recognition errors of few frames. It is noted that a vote is made from the results of the consecutive frames belonging to the same shot.

4 Experiment Results

In out experiments, a total of 172 video clips, approximately 484 minutes, captured from three kinds of sports game are collected to demonstrate the performance of the proposed approach. Other details for video clips are listed in Table 1. The localization of a scoreboard is only performed once for a video of a whole match. Therefore, only selecting short clips is enough for experiments.

Performance evaluation is made on the scoreboard localization and scoreboard text recognition modules separately.

Scoreboard Text Localization: For each video clip, the ground truth of scoreboard bounding box (which mainly contains the score and team information) was

Table 1. Details of tested video clips

Sport type	Frame size	Frame rate(f/s)	Amount	Average Duration(mins)
Soccer	352×288	25	3.4	72
Basketball			1.2	45
Rugby			3.3	55

Table 2. Results of scoreboard text localization

Pixel-based			Text box-based
avmatchrate	avmiss	avfalse	
91.4%	8.6%	7.9%	92.3%

created manually. Two kinds of evaluation for scoreboard localization are tested: pixel-based and text box-based performance numbers.

Pixel-based performance numbers calculate the match rate, miss rate and false rate on the number of pixels the ground truth and the detected scoreboard bounding box have in common (as shown in Figure 6), for detected scoreboard bounding box D_i in i_{th} video clip:

$$matchrate_{pixel-based,i} = \frac{card(D_i \bigcap G_i)}{card(G_i)}$$
$$miss_{pixel-based,i} = 1 - matchrate_{pixel-based,i}$$
$$false_{pixel-based,i} = 1 - \frac{card(D_i \bigcap G_i)}{card(D_i)}$$

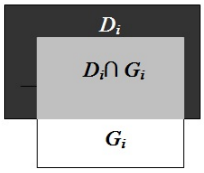

Fig. 6. Diagram of pixel-based evaluation

Here $D_i = \{d_1, d_2, \cdots, d_{n_i}\}$ and $G_i = \{g_1, g_2, \cdots, g_{m_i}\}$ are the sets of pixel set representing the detected scoreboard bounding box and the ground-truth scoreboard bounding box of size n_i and m_i for i_{th} video clip respectively. N is the number of tested video clips. Operator $card\,(\cdot)$ counts the number of elements in a set. The average match rate, average miss rate and average false rate are calculated as follows:

$$avmatchrate = \sum_{i=1}^{N} matchrate_{pixel-based,i}$$
$$avmiss = \sum_{i=1}^{N} miss_{pixel-based,i}$$

$$avfalse = \sum_{i=1}^{N} false_{pixel-based,i}$$

In contrast, the text box-based performance is evaluated by *recall* which refers to the number of detected boxes that match with the ground truth. The created scoreboard text bounding box D_i was regarded as localized correctly if and only if the two boxes D_i and G_i overlapped by at least 85% for i_{th} video clip.

$$recall = \frac{\sum_{i=1}^{N} \delta(D_i, G_i)}{N}$$

Here:

$$y = \begin{cases} 1 & if\ min(ComD, ComG) \geq 0.85 \\ 0 & else \end{cases}$$
$$ComD = card\left(D_i \bigcap G_i\right) / card(D_i)$$
$$ComG = card\left(D_i \bigcap G_i\right) / card(G_i)$$

Experiment results of localization performance are given in Table 2. The localization approach correctly found 92.3% of all scoreboard boxes. And the average match rate can achieve to 91.4% with miss rate 8.6%.

Our experiments show that most of scoreboard text boxes generated by the proposed approach are a little smaller than their corresponding ground-truth text boxes, which leads to the results that the average false rate (7.9%) is relatively small and the average match rate is close to the recall.

Scoreboard text recognition: Scoreboard text recognition is performed as described in section 3.2 on all the correctly localized scoreboards. If the score and team information can be obtained, then we consider the scoreboard is correctly recognized. In our experiments, 88.1%of the correctly localized scoreboards were also recognized correctly. Over all stages, 81.4% ($0.881 \times 0.923 = 0.814$) of all scoreboards were recognized correctly.

5 Conclusions

The scoreboard in sports video is an important semantic clue. Localization and recognition of scoreboards is very meaningful for sports video analysis and processing. According to the observation that scoreboards are attached at fixed locations in the sports video and always exists in all sports video frames, we propose an approach for localizing and recognizing scoreboards based on SIFT points matching. In our experiments on a total of 172 sports video clips, approximately 484 minutes, an average of 91.4% of the scoreboard bounding box are correctly matched with a 7.9% false rate. For localization and recognition, 92.3% of all scoreboards box are correctly localized, and 81.4% of all scoreboards can be recognized.

Furthermore, the localization of a scoreboard is only performed once for a video of an entire match, which is efficient.

In the future, we will extend our study to detect score events of sports games by the recognized scoreboard texts.

Acknowledgment

This paper is supported by the Information Access Disruptions (iAD) Project (Norwegian Research Council), the China Scholarship Council, the National Nature Science Foundation of China (No.60902094) and by Science Foundation Ireland under grant 07/CE/I1147.

References

1. Jain, A.K., Yu, B.: Automatic Text Location in Images and Video Frames, Pattern Recognition. 31(12), pp.315–333 (1998)
2. Ngo, C-W., Chan, Ch.K.: Video Text Detection and Segmentation for Optical Character Recognition, ACM Multimedia Systems. 10(3), pp.261-272(2005)
3. Li, X.J., Wang, W.Q., Jiang, S.Q. Huang, Q.M., Gao, W.: Fast and Effective Text Detection, In: IEEE International Conference on Image Processing (ICIP), pp.969-972 (2008)
4. Chen, D.T., Odobez, M.J., Bourlard, H.: Text Detection and Recognition in Images and Videos, Pattern Recognition. 37(3), pp.595-608 (2004)
5. Lienhart, R., Wernicke, A.: Localizing and Segmenting Text in Images and Videos, In:IEEE Transact. on Circuits and Systems for Video Technology, 12(4), pp.256-268 (2002)
6. Wang, R.R., Wan, J.J., Wu, L.D.: A novel video caption detection approach using Multi-Frame Integration, In:IEEE Proceeding of the 17th International Conference on Pattern Recognition(ICPR), 10(3), pp.449-452 (2004)
7. Tang, X., Gao, X., Liu, J., Zhang, H-Z: A Spatial-temporal Approach for Video Caption Detection and Recognition, In: IEEE Trans. on Neural Networks, 13(4), pp.961-971 (2002)
8. Zhang, D. , Rajendran, R.K., Chang, S-F.: General and Domain-Specific Techniques for Detecting and Recognizing Superimposed Text in Video, In: IEEE International Conference on Image Processing (ICIP), pp.22-25 (2002)
9. Su, Y.M., Hsieh, C.H.: A Novel Model-based Segmentation Approach to Extract Caption Contents on Sports Videos, In: International Conference on Multimedia & Expo .(ICME), pp.1829-1832 (2006)
10. Hsieh, C.H., Huang, C.P, Hung, M.H.: Detection and Recognition of Scoreboard for Baseball Videos, In: International Conference on Intelligent Computing(ICIC), pp.337-346 (2008)
11. Lowe, D.: Distinctive image features from scale-invariant keypoints, Computer Vision. 60(2), pp.91-110 (2004)
12. Ballan, L., Bertini, M., Del Bimbo, A., Serra, G.: Video Event Classification Using Bag of Words and String Kernels, In: International Conference on Image Analysis and Processing, pp.170-178 (2009)
13. Pelleg D., Moore A.: X-means: Extending K-means with efficient estimation of the number of clusters, In:International Conference on Machine Learning, pp.727-734 (2000)
14. Otsu,N.: A Threshold Selection Method from Gray-Level Histograms, In: IEEE Transact. On Systems, Man and Cybernetics, 9(1), pp.62-66 (1979)
15. ABBYY FineReader, http://www.abbyy.com/
16. Gonzalez, R.C., Woods R.E.: Digital Image Processing Second Edition (2007).
17. ReadIRIS, http://www.irislink.com/