ACCV2002: The 5th Asian Conference on Computer Vision, 23--25 January 2002, Melbourne, Australia.

Presence Analysis of Players in Broadcasted American Football Video

Fumi Nishiue Noboru Babaguchi Tadahiro Kitahashi
ISIR, Osaka University
Ibaraki, Osaka, 567-0047, Japan

Abstract

Information about who appears in video is of great importance for its retrieval. In particular, since queries about specific players seem most popular in sports video, their name can be a good index. To obtain such indexes, the first step is to efficiently analyze whether or not the player is present in an image frame. In this paper, we propose a method of presence analysis of players in sports video of American football TV programs. This method focuses on both the spatial arrangement of color regions corresponding to the player's head and body in an image frame and the framing continuity in frame sequences. It aims at finding temporal positions at which the player of a particular team is taken as a close-up or medium shot. The experimental results indicate that this method is effective and efficient.

1. Introduction

Content based indexing is a key technique for retrieval of video[1]. Semantical indexes to video contents that indicate 'who is it?' and 'what is it?' are of great interest. Automated detection of objects from video streams is required so that we can obtain such indexes. The procedure of object detection can be divided into two steps. The first is object presence analysis[2] to find image frames where a target object is present. The second is object segmentation and identification to extract its region from the frame and recognize who or what it is.

In this paper, we propose a method of *presence analysis* of players, imaged as a *medium* or *close-up* shot in broadcasted American football video, which is a prerequisite of automated indexing to focused players. For sports video, we are interested in players or related people like coaches and referees. In particular, queries about the players may be most popular. This suggests that indexing by the player's name is indispensable for the retrieval of sports video. Therefore, efficient and reliable presence analysis of players should be required.

The presence analysis of players is closely related to people detection. So far people detection has been sometimes reduced to face detection. Neural network based methods were proposed by Rowley et al.[3], Sung et al.[4] and Osuna et al.[5]. Satoh et al.[6] applied face detection to video indexing. However, there are some drawbacks in applying them to some kinds of sports video. Most methods are time consuming and need learning process for a large number of training examples. In our application, a player's face is not always visible, because he puts his helmet or cap on.

On the other hand, color information is extensively used in image/video content analysis. For example, color histogram[7], color coherence vectors[8], and color correlogram[9] have been proposed for content based retrieval mainly for still image databases. In addition, extended systems to video applications have been reported in [10, 11, 12, 13]. In this paper, we also focus on the color information characterizing the player of American football games. We want to detect him from his side or back view. To deal with the change of these appearances depending on his pose, size and location, we take account of the *spatial arrangement* of color regions corresponding to his *body* and *head*. In the American football case, a uniform and a helmet stand for them each.

We further concentrate on the feature of video streams as continuous media. In broadcasted sports video, there is a case where a focused player who is a star player or has done a nice play is continuously tracked with a camera for a while. This means that the framing in which the focused player is present does not change largely for some seconds in spite of the camera work like static, pan or zoom-up. We here call this feature *framing continuity*. Considering this, we can avoid detecting other objects that may appear incidentally. This point mainly characterizes our method.

Based on the spatial color arrangements and the framing continuity, the proposed method aims at finding temporal positions at which a player of a particular team is present at a close-up or medium shot with low computational cost. Section 2 describes the method in detail. In Section 3, experimental results and discussions are shown. Section 4 is a conclusion of our work.

2. Presence Analysis of Players

The proposed method consists of three procedure steps.
step1: Registration of color models
step2: Presence analysis for a single image frame
step3: Presence analysis for frame sequences

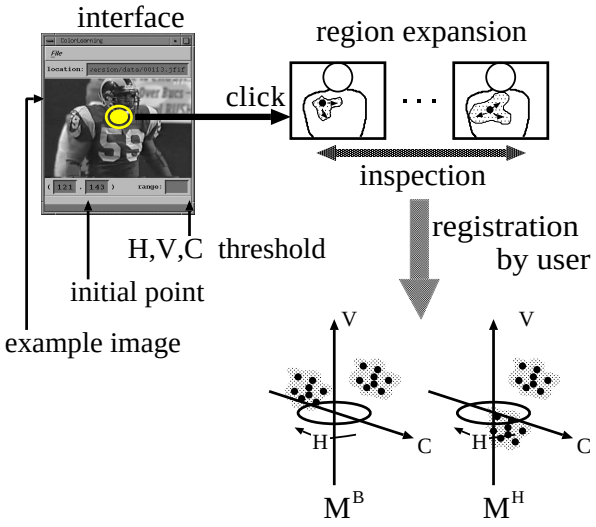


Figure 1. Color model.

We begin by mentioning a couple of symbols and definitions. Let $f_k^u(x,y)$, $(0 \leq x \leq X, 0 \leq y \leq Y, u \in \{R,G,B\})$ denote a color image of the k-th image frame in a video stream. The x and y axes are horizontal and vertical, respectively and the coordinate (0,0) is the left-upper most point. A transformed image $f_k^w(x,y), (w \in \{H,V,C\})$ to the Munsell space is also used. H, V and C denote hue, chroma and value, respectively, where $0 \leq H \leq 99$, $0 \leq V \leq 15$, and $0 \leq C \leq 49$. The transformation from (R,G,B) to (H,V,C) is based on [14].

In this paper, the player to be detected is assumed to be framed as a close-up or medium shot. According to [13], the shot is close-up if the head is visible, whereas it is medium if the body is framed from the waist up. The close-up and medium shots are defined in terms of the ratio R_s of the area $X \cdot Y$ of the image frame to that of a rectangle bounding the head region. If $0.15 < R_s \leq 1$, then the frame is close-up. If $0.05 \leq R_s \leq 0.15$, then it is medium. In what follows, we proceed to describe the procedure steps in detail.

2.1. Registration of color models

A user interface to register color models is provided as shown in Fig.1. First of all, a user selects an example image in which a player of a particular team is present. To make the color model of the body (head), the user points out an initial point (x_0, y_0) in the body (head) region of the player. Once it is determined, the region expansion procedure is operated as follows.

step1: $\mathbf{R}_0 \leftarrow \{(x_0, y_0)\}$
step2: For i = 1,2,3, ...
$\forall (x_{i-1}, y_{i-1}) \in \mathbf{R}_{i-1}$,
$\mathbf{R}_i \leftarrow \mathbf{R}_{i-1} \cup \{(x_i, y_i) \mid |f^w(x_0, y_0) - f^w(x_i, y_i)| \leq \eta_w\}$,
where (x_i, y_i) is an 8-neighboring point of (x_{i-1}, y_{i-1}), $f^w(x_i, y_i)$ is $w(\in \{H,V,C\})$ value of (x_i, y_i), and η_w is a threshold.
step3: if $\mathbf{R}_i = \mathbf{R}_{i-1}$ then step4
otherwise step2
step4: $\mathbf{R} \leftarrow \mathbf{R}_i$

The user can inspect resultant images by changing the initial point and the threshold. When an appropriate region is produced, the color model is constructed as a set of (H,V,C) values in the region $\mathbf{R}$. Each model, denoted by $\mathbf{M}^H$ and $\mathbf{M}^B$, for the head and the body, is constructed from a set of example images consisting of close-up and medium shots. Note that it is common to both shots. Because the model is composed of a set of (H,V,C) values, we can deal with the body or head which has more than two colors. In such a case, each color is registered independently.

2.2. Background elimination

Background elimination is performed as preprocessing. First, we select n points in the image frame. From these points, the region expansion procedure, stated in the previous section, is evoked. Let $\mathbf{B}_k, k = 1, ..., n$ denote an obtained region with the procedure. It is noted that $\mathbf{B}_k$ is a coherent region with respect to the color difference. Pass et al. [8] defined the coherent set as a set of pixels of the same color. Alternatively, our definition allows a little difference of pixel color. We regard the background as the region with almost constant color. The background region $\mathbf{B}$ is formed by merging $\mathbf{B}_k$. A region whose area is small is neglected when it is merged.

Further, we examine the lower portion of the image frame, considering the characteristics of the medium and close-up shots. At both shots, the pixels in the body region are mainly located in the lower portion. If more background regions are located there, there is less possibility that the frame is included in either medium or close-up shot. Thus the following procedure is performed.

For each column of the image frame, we count the number of points included in the lower region of $\mathbf{B}$. The number of such points $r_i, (0 \leq i \leq X)$, is defined as

$$r_i = |\{(i,y) \in \mathbf{B} \mid Y - \lambda Y \leq y \leq Y, (0 \leq \lambda \leq 1)\}|,$$

where X and Y denote the width and height of the image frame, respectively, and λ is a parameter to define the lower portion. In the above equation, the symbol $|\mathbf{S}|$ represents the number of elements in a set $\mathbf{S}$. If r_i exceeds some threshold, all the points on the i-th column are also viewed as the background region. Thus, for $x = 0, ..., X$, the following set

$$\{(x,y) \mid r_x \geq (\lambda Y) \cdot \tau, 0 \leq y \leq Y, (0 \leq \tau \leq 1)\},$$

is added to $\mathbf{B}$. Consequently, in order to pursue efficiency of the processing, a foreground region $\mathbf{D}$ is formed as

$$\mathbf{D} = \mathbf{F_k} - \mathbf{B},$$

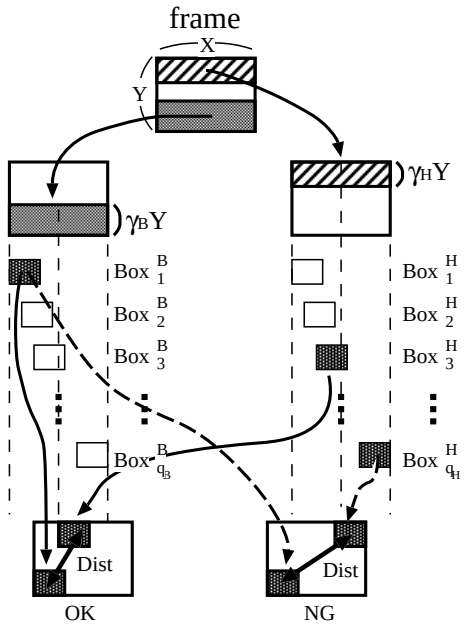


Figure 2. Presence analysis for an image frame.

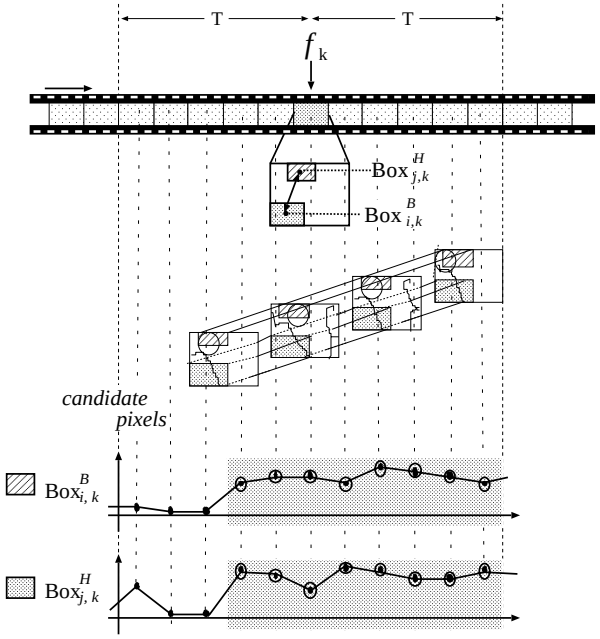


Figure 3. Presence analysis for frame sequences.

where $\mathbf{F_k}$ is the whole region of the image frame.

2.3. Presence analysis for a single image frame

For the foreground region $\mathbf{D}$, we perform presence analysis for a single image frame, considering spatial arrangement of color regions. We call a pixel that meets the color model a *candidate pixel*. From a player's appearances at the close-up and medium shots, the head region and the body one are located in the upper and lower portion of the image frame, respectively. Therefore, the regions in which the head and the body should be detected are restricted in terms of parameters γ_B, γ_H and Y as illustrated in Fig.2.

Next we divide each region into overlapped small regions, called *boxes*, whose number equals to q_B, q_H for the body and the head. Let $\mathbf{Box}_i^B, i = 1, ..., q_B$ and $\mathbf{Box}_j^H, j = 1, ..., q_H$ denote each of the boxes for the body and the head, respectively. Each box is given by

$$\mathbf{Box}_i^B = \{(x,y) \mid \frac{X}{q_B+1}(i-1) \le x < \frac{X}{q_B+1}(i+1), \\ Y - \gamma_B Y \le y \le \gamma_B Y\},$$

$$\mathbf{Box}_j^H = \{(x,y) \mid \frac{X}{q_H+1}(j-1) \le x < \frac{X}{q_H+1}(j+1), \\ 0 \le y \le \gamma_H Y\},$$

where γ_B and γ_H are parameters to define both regions, and X and Y are the width and height of the image frame.

Subsequently, candidate pixels are detected in each box. Let $\mathbf{M}^B$ and $\mathbf{M}^H$ denote each color model for the body and the head, respectively. If

$$\frac{\left|\{(x,y)|\ f^w(x,y) \in \mathbf{M}^B,\ (x,y) \in \mathbf{Box}_i^B \cap \mathbf{D}\}\right|}{|\mathbf{Box}_i^B|} \ge \rho_B,$$

$$\frac{\left|\{(x,y)|\ f^w(x,y) \in \mathbf{M}^H,\ (x,y) \in \mathbf{Box}_i^H \cap \mathbf{D}\}\right|}{|\mathbf{Box}_j^H|} \ge \rho_H,$$

then it is determined that there are color regions matching the color model in the box. These boxes are called *matched boxes*.

The positional relation between matched boxes allows us to decide the player's presence. Specifically, if all the distances between the matched $\mathbf{Box}_i^B$ and $\mathbf{Box}_j^H$ exceed a constant value, it is determined that no player is present in the image frame. Fig.2 shows the behavior of this process.

2.4. Presence analysis for frame sequences

We make use of the framing continuity for presence analysis for frame sequences. The pair of matched boxes for the head and the body appears continuously while a player is focused in the video stream. This framing continuity prevents us from detecting incidental presence of the players that are not focused on.

Fig.3 is a rough sketch of presence analysis for frame sequences. Now let f_k be an image frame of concern. The two

curves represent the numbers of candidate pixels in $\mathbf{Box}_{i,k}^{B}$ and $\mathbf{Box}_{j,k}^{H}$. The shaded portion is the time interval when the positional relation between the two boxes is kept unchanged.

Consider T frames before and after the frame f_k. We form three sequences as

$$(f_{k-T}, ..., f_k), (f_{k-T/2}, ..., f_{k+T/2}), (f_k, ..., f_{k+T}).$$

Note that there are T frames in each sequence. We check whether or not each frame in the sequence has the simultaneously matched boxes where the positional relation between the head and the body is maintained. For each of the three sequences, if

$$\frac{\#\ \text{frames with simultaneously matched boxes}}{T}$$

exceeds a threshold ξ, then it is determined that a player is present in the frame f_k, and this is referred to as positive presence.

3. Experimental Results

3.1. Setup

The method was implemented with the C language on SGI Octane (MIPS R12000 300MHz). The input video rate was six frames per second. The image frame was of size 160(X) $\times$ 120(Y) with 256-level RGB values.

To evaluate accuracy of the proposed method, we carried out some experiments. We employed three sample streams, whose details are shown in Table 1. All the streams were actual broadcasted American football programs. We have to take account of a variety of colors so that we can evaluate the performance from diverse viewpoints. The colors are classified into either chromatic colors such as red, blue and green or achromatic ones such as white, grey and black. In this case, the seven chromatic colors and the five achromatic ones were considered.

To construct each of the color models, four image frames consisting of both close-up and medium shots were extracted from each sample stream. The time interval in the stream where the color model was acquired was from zero to ten minutes. The method was tested for the stream of ten minutes except for the interval where the color model was acquired. The target objects to be detected were the players of the offense team because they more frequently appeared in the broadcast than those of the defense team. Each interval for the test was selected in order to get the sufficient targets.

We evaluated the accuracy for each shot. If six consecutive image frames of positive presence are found, then we determine the player is detected in the shot. It implies that he is present for one second in the video.

As evaluation measurement, the recall and precision were introduced. As is well known, they have trade-off relation, so F-value was also considered for overall evaluation.

Table 2. Results of presence analysis.

Team	Recall	Precision	F-value
KG	92.3%(24/26)	88.9%(24/27)	90.6
HS	100%(18/18)	75.0%(18/24)	85.7
GB	92.3%(12/13)	80.0%(12/15)	85.7
NE	93.3%(14/15)	66.7%(14/21)	77.8
DEN	73.3%(11/15)	31.4%(11/35)	44.0
KC	94.7%(18/19)	85.7%(18/21)	90.0
Total	91.5%(97/106)	67.8%(97/143)	77.9

Table 3. Recall rates for close-up and medium shots.

Team	Close-up	Medium
KG	90.0%(18/20)	100%(6/6)
HS	100%(14/14)	100%(4/4)
GB	100%(4/4)	88.9%(8/9)
NE	90.0%(9/10)	100%(5/5)
DEN	100%(2/2)	69.2%(9/13)
KC	100%(12/12)	85.7%(6/7)
Total	95.2%(59/62)	86.4%(38/44)

Their definition is as follows.

$$\begin{aligned}
\text{Recall} &= \frac{\#\ \text{correctly detected shots}}{\#\ \text{actual presence shots}} \times 100 \\
\text{Precision} &= \frac{\#\ \text{correctly detected shots}}{\#\ \text{all detected shots}} \times 100 \\
\text{F-value} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}$$

A set of parameters was given as $\lambda = 0.15, \tau = 0.85, \gamma_B = 0.35, q_B = 3, \gamma_H = 0.45, q_H = 3, \rho_B = 0.15, \rho_H = 0.03, T = 9$, and $\xi = 0.8$ throughout the experiment. These were decided empirically.

3.2. Results of presence analysis

Using the sample streams of video-A, B and C, we analyzed the presence of players of the six teams, i.e. KG, HS, GB, NE, DEN and KC (cf. Table 1). Table 2 shows the recall, precision and F-value for each team. In addition, Table 3 summarizes the recall rates for close-up and medium shots. Examples of correct detection are shown in Fig. 4, where Fig. 4(a)(b) and (c)(d) depict close-up and medium shots, respectively. Examples of false detection are shown in Fig. 5.

As shown in Fig. 4, it is possible to detect the players in various appearances. The players from their side or back view can be detected. More reliable detection was attained for close-up shots. In these shots, the head region which is,

Table 1. Sample streams.

Sample	Game	Team	Color (Chromatic/ Achromatic)	
			Head	Body
video-A	Koshien Bowl 2000(Japan)	Kangaku Univ.(KG)	Light Blue (C)	Light Blue (C)
		Housei Univ. (HS)	Orange (C)	White (A)
video-B	Super Bowl 1997	Green Bay Packers(GB)	Yellow (C)	Green (C)
		New England Patoriots(NE)	Silver (A)	White (A)
video-C	NFL2000	Denver Broncos(DEN)	Navy (A)	White (A)
		Kansas City Chiefs(KC)	Red (C)	Red (C)

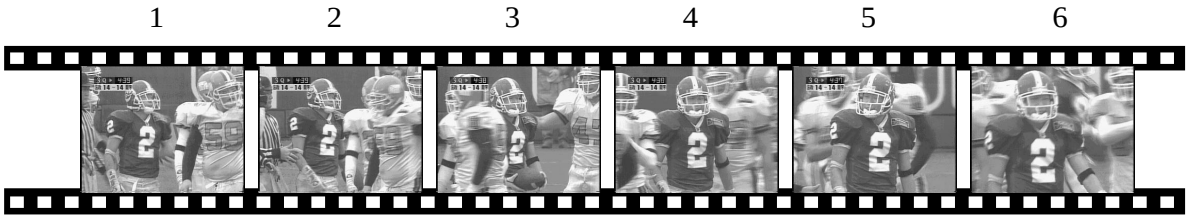


Figure 6. Example of incidental presence.

in fact, represented as the helmet is relatively dominant. We guess that the color features in the head region might keep stable.

Let us now think about false and miss detection. First, false detection was likely to take place in the image frame where a non-player like a coach wearing clothes in the color similar to the players was present, as shown in Fig.5(a). On the other hand, some long shots, at which the whole figure is visible in the image frame, were erroneously detected as medium shots, as shown in Fig.5(b). Consequently, false detection was observed in the image frames where there existed similar color arrangements. Because there were a lot of white regions in the background, they were sometimes erroneously detected as the white uniforms.

Miss detection occurred in case the body color was largely changed because of the change of lighting environment. If the body color was achromatic, namely white or silver, we strongly suffered from the lighting influence. For example, the whitish uniforms turn dark grey in the shade. For this lighting problem, a compensation method may be promising by considering the overall brightness of the image frame.

As a result, this method is disadvantageous to achromatic colors. From Table 2, we notice that the lowest two teams with respect to their accuracy were NE and DEN, each of which has the white uniforms. Due to a lot of false detection, the precision rate considerably decreased. If we exclude DEN and NE from the evaluation, the recall and precision will rise to 94.7% and 82.7%, respectively.

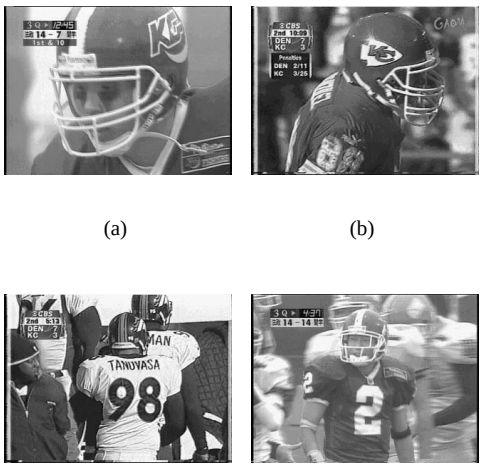
We investigated the possibility to exclude incidental presence of other players. By considering the framing continuity, this method tries to exclude the incidental presence of other players that are not focused on. In other words, it tries to detect the players selectively. To examine this, we tested this method for just the same interval of video-A, considering two cases: 1) detecting offense players, denoted by KG, and 2) detecting defense ones denoted by HS.

Table 4. Presence analysis for the same interval.

Team	Recall	Precision	F-value
KG	92.3%(24/26)	88.9%(24/27)	90.6
HS	100%(4/4)	50.0%(4/8)	66.7

Table 4 shows the experimental result. The target objects for the defense team HS appeared less frequently. A good example of selective detection is shown in Fig.6. For this shot, only the offense player who is numbered 2 was detected in the case 1). Conversely, as you can see at the 2nd and 3rd frames, another player with a white uniform is going across. He was not really detected in the case 2) though his figure was in the range of medium shots. These results demonstrate that selective detection is possible, and the framing continuity produces a good effect on presence analysis of the focused players.

At present, this method is based only on the enumeration of the pixels matching the color model, because we have concentrated on its simplicity. However, we have to take account of the shape of the color regions for improvement.



(a) (b)

(c) (d)

Figure 4. Examples of correct detection.

(a) (b)

Figure 5. Examples of false detection.

In other words, the likeliness about each shape of the head and the body should be considered.

Finally, we discuss the efficiency of this method. In our implementation, the processing time was 0.101 sec per an image frame. Since the input frame rate is currently six fps, the processing is real time, but requires three times speed-up to work at the video rate, 1/30 sec. For comparison, we measured the processing time of the existing face detector[3], which is often applied to video indexing. For an image frame of the same size, the processing time was 0.360 sec, which is about three times as long as ours.

4. Conclusion

We addressed a simple method for presence analysis of players in broadcasted American football video. From the experimental results, we obtained the recall and precision of 94.7% and 82.7% for detection of the players in chromatic colors. The results indicate that the introduced two features, the spatial arrangements of color regions and the framing continuity, are effective for the task. Although this method was analyzed only for the video of American football, it may be applicable to many kinds of sports video because it is designed by taking general features and knowledge into consideration. The remaining work is to improve the performance on objects in achromatic colors, to apply this method to other kinds of video, and to develop a method for player identification.

Acknowledgments–We thank Dr. Rowley and Prof. Satoh for providing us with the face detector program. This work is partly supported by a Grant-in-Aid for scientific research from the Japan Society for the Promotion of Science and by Telecommunications Advancement Organization of Japan.

References

[1] A. Del Bimbo, "Visual Information Retrieval," Morgan Kaufmann, 1999.

[2] P. Aigrain, H. J. Zhang, and D. Petkovic, "Content–Based Representation and Retrieval of Visual Media, A State–of–the Art Review," *Multimedia Tools and Applications*, 3, pp.179–202, 1996.

[3] H. A. Rowley, S. Baluja and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. on PAMI*, Vol.20, No.1, pp.23-38, 1998.

[4] K.-K. Sung and T. Poggio, "Example-Based Learning for View-Based Human Face Detection," *IEEE Trans. on PAMI*, Vol.20, No.1, pp.39-51, 1998.

[5] E. Osuna, R. Freund and F. Girosi, "Training Support Vector Machines: an Application to Face Detection," *Proc. CVPR97*, pp.130-136, June, 1997.

[6] S. Satoh, Y. Nakamura and T. Kanade, "Name-It: Naming and Detecting Faces in News Videos," *IEEE Multimedia*, Vol.6, No.1, pp.22-35, 1999.

[7] M.J. Swain and D.H. Ballard, "Color Indexing," *Intl. J. Computer Vision*, Vol.7, No.1, pp.11-32, 1991.

[8] G. Pass, R. Zabih and J. Miller, "Comparing Images Using Color Coherence Vectors," *Proc. ACM Multimedia 96*, pp. 65–73, 1996.

[9] J. Huang, S.R. Kumar, M. Mitra, W-J. Zhu and R. Zabih "Image Indexing Using Color Correlograms," *Proc. CVPR97*, pp.762-768, 1997.

[10] A. Pentland, R. Picard and S. Sclaroff, "Photobook: Tools for Content-Base Manipulation of Image Databases," *Intl Journal of Computer Vision*, Vol. 18, No. 3.,pp. 233-254, 1996.

[11] M. Flickner, et al., Query by Image and Video Content: The QBIC System, *IEEE Computer*, pp.23-32, 1995.

[12] J. R. Smith and S.-F. Chang, "VisualSEEk: a Fully Automated Content-based Image Query System," *Proc. ACM Multimedia 96*, Nov., 1996.

[13] R. Lienhart, W. Effelsberg and R. Jain, "VisualGREP:A Systematic Method to Compare and Retrieve Video Sequences," *Multimedia Tools and Applications*, Vol.10, pp.47-72, 2000.

[14] M. Miyahara and Y. Yoshida, "Mathematical Transform of RGB Color Data to Munsell HVC Color Data," *SPIE Visual Communication and Image Processing'88*, 1001, pp.650-657, 1988.