コンピュータビジョン 97－7
（1995．11．16）

Focussed Color Intersection for Content Based Retrieval

V V Vinod vinod@apollo3.brl.ntt.jp
Chic Hashizume hasizume@eye.brl.ntt.jp
Hiroshi Murase murase@apollo3.brl.ntt.jp

NTT Basic Research Labs
3-1, Morinosato Wakamiya, Atsugi-Shi
243-01 JAPAN

In this paper we propose a focussed color intersection strategy for extracting known objects in a complex scene. The method utilizes only the color distribution of the objects and the scene. It has applications in object extraction, content based retrieval and tracking are presented.

局所的色交差法による Content Based Retrieval

V V Vinod　橋爪千枝　村瀬洋

NTT基礎研究所

本報告は色の情報を利用して複雑な背景を持つ画像から特定の物体を抽出する手法について述べたものである．本手法では，画像の各部分について色ヒストグラムを局所的に計算し，モデル物体の色ヒストグラムと照合することにより，その物体がその画像中に存在するかどうか，また存在する場合にはどこに存在するかを検出する．本手法を複雑背景からの物体の抽出，コンテントに基づく検索，物体のトラッキングに適用し，その有効性を示す．

1 Introduction

Content based image retrieval may be defined as the process of retrieving all scenes which contain a desired image. Such retrieval is central to any image or multimedia database system [8]. The major task here is to identify the known objects in an image. We present a strategy for detecting the presence of known objects and extracting the approximate region it occupies in the scenes using only color information. This task has applications in scene interpretation, content based retrieval from large image databases, tracking objects in image sequences etc.

Most of the strategies proposed for identifying and locating objects in a scene utilizes geometric features [3, 10]. Such methods are expensive and are adversely affected by changes in the scene which would result in a change in the perceived shape. Moreover, they cannot be applied to non-rigid objects. Template matching techniques constitute the other approach for detecting objects in a scene under small distortions and noise [7, 10, 6]. Storing and matching against several templates for the same model could be computationally very expensive. Hence it would be desirable to do the matching using features invariant to as many changes as possible. The color distribution of objects present one such feature and is adopted in this work.

Recently it has been shown that color distributions can be efficiently used for image matching. Swain and Ballard [11] introduced an efficient technique called Histogram Intersection which evaluates the similarity between two color images. This technique has been proposed for indexing into large image databases. Modifications and enhancements to Histogram Intersection and other techniques may be found in [2, 4] and [5]. A color based image retrieval system using fuzzy matching techniques has been proposed in [1]. Schettini [9] applied color matching along with shape matching for detecting a known object against a known background.

The above approaches indicate that color constitutes an important feature for image matching. However, all the color histogram based methods except that in [9] confine themselves to evaluating the similarity between two images. This limits the role to indexing image databases given most part of the image and cannot be applied when the retrieval has to be based on a small portion of the image. In such cases, matching the histogram of the scene with that of the model with fail to provide much information. Hence an appropriate color histogram matching strategy which focuses its search on different regions of a complex scene is expected to perform well in the task of content based retrieval.

We propose an iterative color matching strategy which focuses on parts of a scene. The histogram intersection technique is adopted for evaluating the match between a focus region and a model. The method is stable against changes in 2-dimensional orientation, some amount of occlusion and moderate changes in shape. The method, however, is influenced by changes in lighting conditions and major changes in 3-dimensional orientation.

The focussed color intersection method is given in section 2. Experimental results are given in section 3, and conclusions in section 4.

2 Focussed Color Intersection

The following general setting is considered for developing the method. "Given a set of models $\mathsf{M} = \{M_n\}$, $n = 1, \ldots, N$ where each M_n is the color image of a known object and an input scene $\mathcal{I}$ of $X \times Y$ pixels (i.e., $\mathcal{I} = \mathrm{p}_{xy}$, $x = 1, \ldots, X$, $y = 1, \ldots, Y$), identify any model objects present in the scene and extract the regions occupied by them". The input scene $\mathcal{I}$ may consist of zero or more known objects against a complex unknown background. The sizes of the objects may vary across scenes. There could be any amount of change in 2 dimensional orientation and small changes in 3 dimensional orientations. Objects may be partially occluded and its shape may vary from scene to scene. The method consists of three steps - focus region identification, iterative matching and pruning and confidence evaluation. Each of these steps are described below.

2.1 Focus Regions Identification

Since more that one object may be present in the scene, we have to focus on parts of the input scene for matching against the models. Since the size of the objects may vary from scene to scene, the focus regions should span small parts of the scene as well as comprise of the entire scene. However, since the color distribution of very small regions in the scene will not carry any effective information, the focus regions should cover a sufficiently large part of

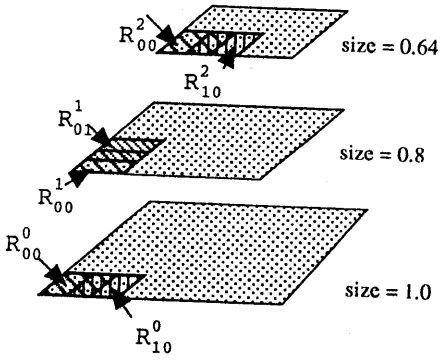

Figure 1: The focus region extraction process

the input scene. The set of focus regions are derived as given below.

In the absence of a priori information favoring any particular shape for the different regions to be considered, a regular shape such as a circle or square may be used. For the sake of concreteness, we consider a square shape. Focus regions are extracted by scanning the image with a square window of size $w \times w$ pixels. For scanning an image the window is shifted by s pixels in one direction at a time. After one complete scan is over, the input image is resized by scaling it by a factor α where $\alpha < 1$. Focus regions are extracted from this resized image by scanning with the same window as earlier. By this process we would be extracting larger regions from the original image to focus upon. This process of resizing by a factor of α and scanning the input image is continued until the image becomes smaller than the scanning window. Figure 1 depicts this process. The original image (size = 1.0) and two resized images (sizes 0.8 and 0.64) are shown in the image. The complete set of focus regions extracted may be characterized as follows. Let $\mathcal{I}^k$ denote the image resized by α^k and $\mathbf{p}^k_{xy}$ denote the pixels belonging to $\mathcal{I}^k$. Then

$$\mathcal{I}^k = \mathbf{p}^k_{xy} \quad x = 1, \ldots, \alpha X \quad y = 1, \ldots, \alpha Y$$
$$\text{where} \mathbf{p}^k_{xy} = \mathbf{p}_{uv}, \quad u = \left\lfloor \frac{x}{\alpha^k} \right\rfloor \quad v = \left\lfloor \frac{y}{\alpha^k} \right\rfloor$$

Let R^k_{ij} denote a focus region belonging to $\mathcal{I}^k$. Then the set R of all focus regions considering all resized images is given by

$$\begin{aligned} \mathsf{R} &= \{R^k_{ij}\} \quad \text{where} \qquad (1) \\ k &= 0, \ldots, \min\left(\left\lfloor \log_\alpha \frac{w}{X} \right\rfloor, \left\lfloor \log_\alpha \frac{w}{Y} \right\rfloor\right) \\ i &= 0, \ldots, \frac{\alpha^k X - w}{s}, j = 0, \ldots, \frac{\alpha^k Y - w}{s} \\ R^k_{ij} &= \mathbf{p}^k_{xy} \quad x = si+1, \ldots si+w \\ &\quad \text{and } y = sj+1, \ldots sj+w \\ &\quad \mathbf{p}^k_{xy} \text{ is not masked} \qquad (2) \end{aligned}$$

Initially none of the pixels are masked. Subsequently when a focus region is associated to a model, pixels belonging to that region are masked. The focus regions R derived as above and the set of models M constitute the initial sets R_c and M_c of competing focus regions and models respectively.

2.2 Matching and Pruning

Each region in the set of focus regions R_c is matched against each of the models in M_c. The quality of the match is evaluated using the histogram intersection technique proposed by Swain and Ballard [11]. The 3-dimensional histogram of each model is constructed by taking into account all pixels in the model's image. The histogram counts are divided by the total number of pixels in the model's image. This normalized histogram constitutes the representation of the model. The histogram of the focus regions are constructed using the same quantization of the color space as that used for model histograms. While constructing the histogram of a region any pixels which are masked are not considered. The focus regions histogram counts are also normalized by dividing by the total number of pixels considered for constructing the histogram. From the normalized histograms, the histogram intersection value $I(M, R)$ for each competing model $M \in \mathsf{M}_c$ and focus region $R \in \mathsf{R}_c$ is computed as
$I(M, R) = \sum \min$(normalized histogram count of M, normalized histogram count of R) where the sum is taken over all cells of the histogram.

In the case of a perfect match between M and R the histogram intersection value $I(M, R)$ will be equal to 1.0 which is very unlikely. In general, even when R contains exactly the same object as M, the intersection value would be less than 1.0. At the same time very low values of $I(M, R)$ may be caused due to partial similarity between models and/or background pixels and other noise. They do not indicate the presence of the model object. We eliminate all matches with very low values by applying a low threshold θ. Now, several models may have intersection value above the

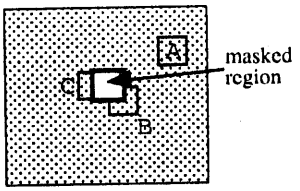

Figure 2: The effect of masking pixels in a focus region on other focus regions in the image

threshold θ for the same or overlapping focus regions. If several models are competing for the same set of pixels then there should be only one winner emerging. However, if there are more than one disjoint set of pixels for which the models are competing then more than one model may finally emerge as winners. That is, the best region-model matches such that the regions are disjoint have to be detected. We proceed as follows.

Let the pair (M', R') have the highest intersection value among all the model region pairs. i.e., $I(M', R') = \max_{M \in \mathsf{M}, R \in \mathsf{R}} I(M, R)$ Then M' has the maximum evidence for being present in R' and M' is accepted as the winner. The pixels belonging to R' are masked to prevent them from being matched against other models. The effect of this masking is schematically shown in figure 2. The masked region and three other regions are shown in the figure. The region 'A' has no pixels in common with the masked region and hence remains unchanged. On the other hand regions 'B' and 'C' overlap the masked region and get modified. The region 'C' has few unmasked pixels and its color distribution will not, in general, constitute a good feature. Now, the effect of masking is not restricted to a given image size but will prevail across all resized images. This is depicted in figure 3. In general, regions with a large number of masked pixels do not carry much information for color matching. Hence all regions for which the fraction of unmasked pixels is less than some constant $\beta < 1$ are removed from the set of competing regions. The new set of competing regions R'_c becomes

$$R'_c = \{R \text{ such that } R \in R_c \text{ and fraction of unmasked pixels in } R \text{ is greater than } \beta\} \quad (3)$$

The set of competing regions are pruned in this way after every match step. It may be noted that since at least the region R' is removed from the set of competing focus regions, this

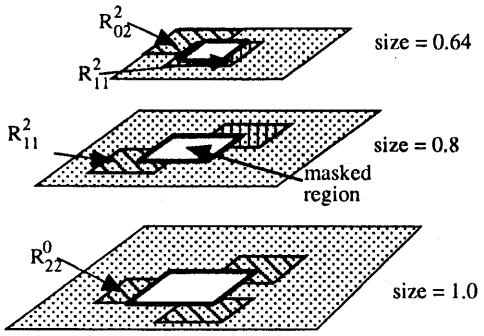

Figure 3: The effect of masking pixels in a focus region of image size 0.8 on other focus regions

set strictly decreases after every match and prune step.

It may not be necessary to consider all the models for matching in later steps. That is the set of competing models may also be pruned. Consider a model M such that $\max_{R \in \mathsf{R}} I(M, R) < \theta$. The match confidence for this model M can increase beyond θ in the next match cycle only due to the influence of masking pixels belonging to the region selected as the present winner. If the number of pixels masked in any region in the present step is low then the intersection value will not change much. On the other hand if a large number of pixels are masked such regions do not carry much information for color matching. Consequently, the maximum match confidence of any model is not expected to increase much in later match steps. Specifically models for which the maximum match value quite less than θ have very little likelihood of emerging as winners in later cycles and may be pruned. We prune the set of competing models as

$$M'_c = \{M \text{ such that } M \in \mathsf{M}_c \text{ and } \max_{R \in \mathsf{R}} I(M, R) \geq \theta\} \quad (4)$$

It may be mentioned that pruning M_c increases the efficiency of the matching process and it is not necessary for the working of the method. Also, M_c may be pruned using a cut-off value lower than θ so that more models take part in the matching process.

Thus in each match and prune step one region is associated to a model and the set of focussed regions as well as the set of competing models are pruned. If the pruned set of focus

regions R'_c and the pruned set of competing models M'_c are not empty then the match and prune process continues by matching the regions in R'_c against the models in M'_c. By this process, eventually the set of competing focus regions and/or the set of competing models will become empty. Then the iterative process of matching and pruning terminates with a set of regions associated to those models which had emerged as winners in some match and prune step.

2.3 Confidence Evaluation

Focus region identification, followed by matching and pruning will yield a set of regions associated with each model which is identified as present in the scene as well as a match confidence for each region. However, it may be observed that the individual match confidences of the regions associated with a model are only indicative and not an absolute measure for the presence of the model. For example the match value between the union of two disjoint regions belonging to a partially occluded object and the model may be higher than that of either of the regions considered individually. Similarly individual regions may have high confidence values but put together they may not resemble any model. Thus, the confidence measure will have to consider the totality of all regions associated with an object. Once the match and prune cycle is over a combined histogram of all the regions associated with a model is computed and its intersection with the model histogram is evaluated. This histogram intersection value serves as absolute match confidence for a particular object.

The algorithm for focussed color intersection is specified below, followed by the experimental results in section 3.

1 $\mathsf{M}_c = \mathsf{M}$ and $\mathsf{R}_c = \mathsf{R}$ where R is defined by equation 1 and M is the set of models.

2 Compute the color histogram intersection value $I(M, R)$ for all $R \in \mathsf{R}_c$ and $M \in \mathsf{M}_c$.

3 Let $I(M', R') = \max_{M \in \mathsf{M}, R \in \mathsf{R}} I(M, R)$. Associate region R' with model M'

4 Mask all pixels belonging to focus region R'. Modify all focus regions accordingly.

5 Evaluate the pruned set of regions R'_c and the pruned set of models M'_c following equations 3 and 4 respectively.

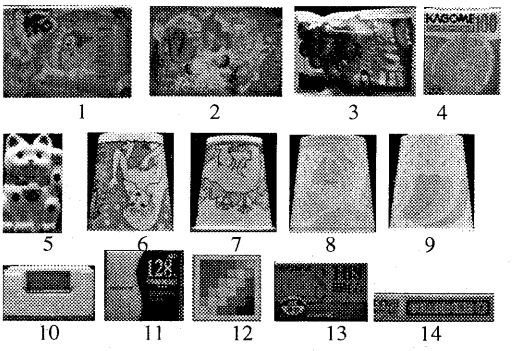

Figure 4: The set of models

6 If $\mathsf{M}'_c \neq \emptyset$ and $\mathsf{R}'_c \neq \emptyset$ then assign $\mathsf{M}_c = \mathsf{M}'_c$, $\mathsf{R}_c = \mathsf{R}'_c$ and go to step 2.

7 For each model M having at least one region associated with it compute $I(M, R_1 \bigcup R_2 \bigcup \ldots \mathsf{R}_m)$ where $R_1, \ldots, R_m$ are the regions associated with M.

3 Experimental Results

Experimental results for object extraction, content based retrieval and tracking objects in image sequences are presented below. The fourteen models used for experiments in 3.1 and 3.2are shown in figure 4. The intensity (I), hue (H) and saturation (S) histograms were used for matching in sections 3.1 and 3.2. Each model was represented by its normalized IHS histogram. The scenes consisted of one to six model objects as well as other objects. The image of each scene was scaled to 128x128 pixels. The parameters α, β and θ were fixed at $\alpha = 0.8$, $\beta = 0.4$ and $\theta = 0.3$.

3.1 Object Extraction

Figure 5 show the results obtained for 6 images. The histograms were constructed using 5, 50 and 40 divisions along the I, H and S axes respectively. A square window of 32x32 pixels was shifted by four pixels at a time for scanning the images. In figure 5 the images are labeled 1 to 6. The objects extracted from scenes 1 to 4 are given in the column below the scene. It may be seen that all model objects present are correctly detected and extracted. Scenes 5 and 6 contain no objects and none was detected. From the results, it may

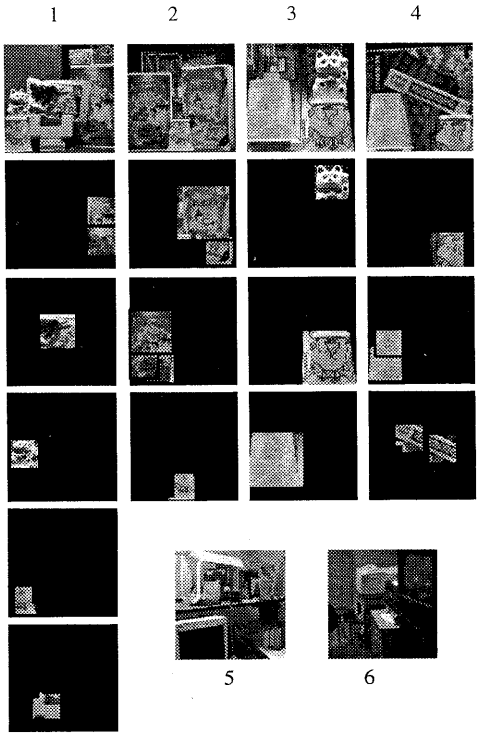

Figure 5: Scenes and extracted objects

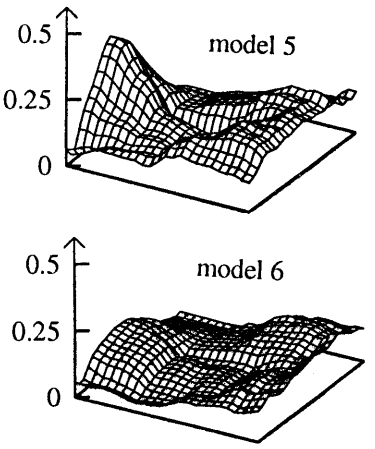

Figure 6: The distribution of histogram intersection values

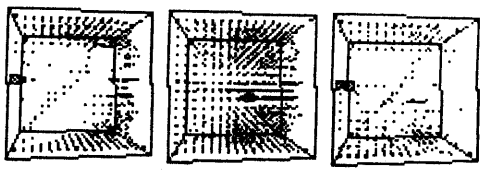

Figure 7: The histogram of (a) model, (b) scene (c) focus region containing model

be observed that the method is stable against changes in the background, orientation, shape and size.

The intersection values for focus regions in the 128x128 pixel image of the scene numbered 4 with models 5 and 6 are shown in figure 6. The distribution for model 5 which is present in the scene has a distinct peak and the distribution for model 6 which is not present in the scene does not have any clear peak. This shows that, the color histogram intersection between models and focus regions constitute an efficient discriminant for detecting and locating objects. In figure 7 the 3-dimensional I-H-S histograms of model number 5 (marked (a)), scene number 4 (marked (b)) and the focus region of scene 4 containing model 5 (marked (c)) are shown. In figure 7 the size of the black boxes are proportional to the histogram values. From this figure it may be observed that, the focus region's histogram (c) is quite similar to the model histogram (a) whereas the entire scenes histogram (b) is quite different. Consequently, histogram intersection per se is not sufficient when the model constitutes only a part of the scene, focusing on parts of the scene is a must. Studies were conducted by varying the histogram bin size, scanning window size and the number of pixels by which the window is shifted for scanning the input image. From the results it was observed that the method correctly identifies the objects under moderate variations in these parameters. The statistics are not presented here due to paucity of space.

3.2 Content Based Retrieval

Here, we first consider the problem of retrieving stored scenes which contain at least one instance of a given object. The method described in section 2 is applied to all the stored scenes with the given object as the only model. All parameters are kept the same as in the case of object extraction. Images were retrieved from a set of 40 stored images. A total of 14 retrievals were done using one model in figure 4 for each retreival. The number of times a scene containing the object was missed and the number of times one, two or more

Number of times			
missed scenes	false retrievals		
≥ 1	$= 1$	$= 2$	> 2
0	3	1	1

Table 1: Results of content based retrieval

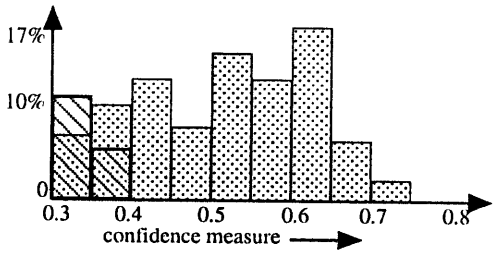

Figure 8: Percentage of retrievals against confidence measure. Hatched columns represent false retrievals

than two scenes containing the given object was falsely retrieved are given in table 1. In no case was a a scene containing the object missed. However, for models 2, 5 and 8 one scene was falsely retrieved. And for models 11 and 4 there were two and four false retrievals respectively. The distribution of confidence measures for false and correct retrievals are shown in figure 8. It may be observed that the confidence measure for false retrievals are close to θ. Since a significant number of correct retrievals also have confidence close to θ, increasing θ will not suffice. However, in cases where false retrievals are a critical, scenes with confidence measures close to θ may be further processed. This processing would be considerably easy since the object regions are already extracted.

In the second experiment the aim was to retrieve all frames containing 'doraemon' from a set of 150 frames of a popular japanese animation movie. The frames were extracted from a $2\frac{1}{2}$ minute sequence at the rate of 1 frame per second. A representative frame and the model are shown in figure 9. Out of 43 frames containing 'doraemon' only two were missed. However, there were 14 false retrievals. Stricter confidence measure evaluation would be required to reduce the false retrievals.

3.3 Tracking objects in image sequences

The method presented in section 2 was slightly modified and applied to the problem of tracking an object in a sequence of images. The RGB histogram was used for matching and the image resizing was restricted to sizes 0.9, 0.95, 1.0, 1.05 and 1.1. This was adopted since large changes in size are not expected across consecutive frames. A dynamic model strategy was adopted to account for variations across different frames. The RGB histogram of a rectangular window in the initial frame (frame 0) containing the object(s) to be searched for is taken as the initial model H_0. Subsequently the model is updated as follows:

$$\begin{aligned} H_1 &= H_0 \\ H_{i+1} &= (1-\lambda_1-\lambda_0)H_i' + \lambda_0 H_0 + \lambda_1 H_i \end{aligned}$$

where H_i is the model histogram used to search for the object in frame i, and H_i' is the histogram of the search result in frame i. The above equation denotes weighted addition of each cell of the histograms. The results obtained for two sequence of frames are shown in figure 10. The value of λ_0 and λ_1 in the model updating equation were chosen as 0.1 and 0.8 respectively. It may be mentioned that in experiments conducted over several sequences of around 250 frames (around 20 seconds at 12 frames per second) the object of interest was correctly tracked in all frames.

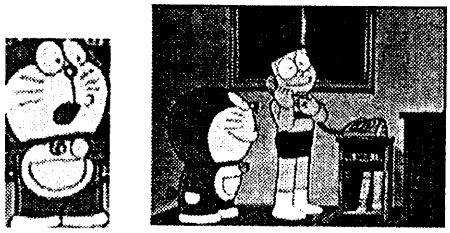

Figure 9: Representative frame and model

4 Conclusion

We have presented a focussed color intersection method for identifying and extracting known objects from a complex scene using only color distributions. Experimental results for object extraction, content based retrieval and tracking objects in image sequences have been presented. From these results we see that focussed color intersection works well for all these applications. For improved accuracy, other features, local as well as global, may be taken into account during the confidence measure evaluation stage. Techniques

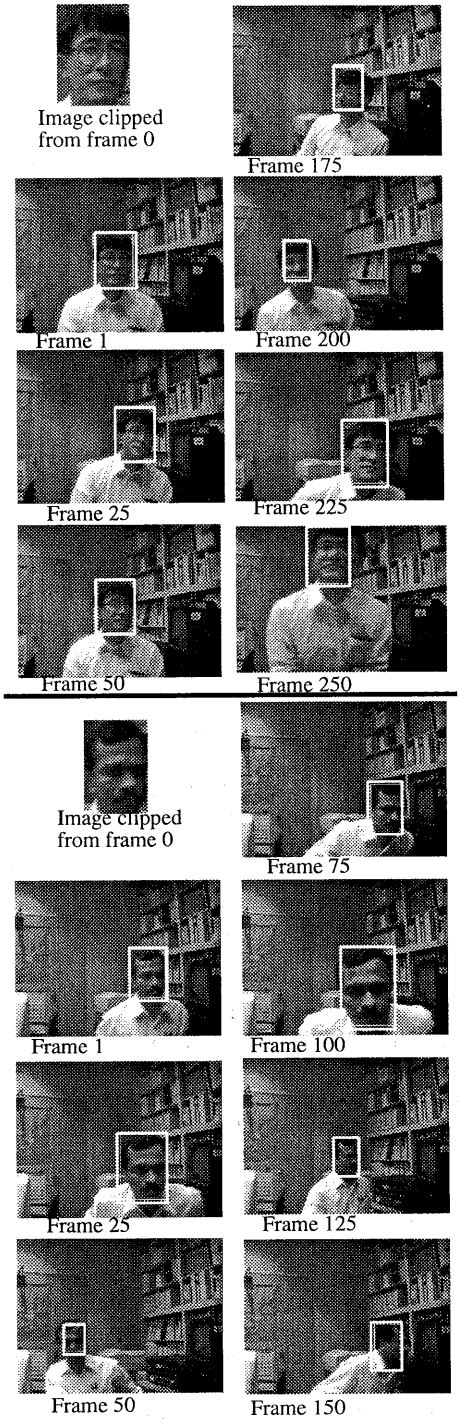

Figure 10: Results of Tracking

for improved confidence measure evaluation and more robust color matching techniques are under investigation.

Acknowledgements The authors wish to thank Dr. T.Ikegami, Dr. K.Ishii, Dr. N.Hagita and Dr. S.Naito of NTT Basic Research Labs for their help and encouragement in conducting this research. The first author thanks NTT Basic Research Labs for the opportunity to conduct research there.

References

[1] Binaghi E, I Gagliardi, and R Schettini. Image retrieval using fuzzy evaluation of color similarity. *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 8, No. 7, pp. 945–967, August 1994.

[2] B V Funt and G D Finlayson. Color constant color indexing. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 17, No. 5, pp. 522–529, 1995.

[3] W E L Grimson. *Object Recognition by Computer: The Role of Geometric Constraints.* The MIT Press, 1990.

[4] J Hafner, H S Sawhney, W Equitz, M Flickner, and W Niblack. Efficient color histogram indexing for quadratic form distance functions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 17, No. 7, pp. 729–736, 1995.

[5] B M Mehtre, M S Kankanhalli, A D Narasimhalu, and G C Man. Color matching for image retrieval. *Pattern Recognition Letters*, Vol. 16, pp. 325–331, 1995.

[6] H Murase and S K Nayar. Image spotting of 3d objects using parametric eigenspace representation. In *Proceedings of the 9th Scandinavian Conference on Image Analysis*, June 1995.

[7] A Rosenfeld and A C Kak. *Digital Picture Processing.* Academic Press, 1976.

[8] M Sakauchi. Database vision and image retrieval. *IEEE Multimedia*, pp. 79–81, Spring 1994.

[9] R Schettini. Multicolored object recognition and location. *Pattern Recognition Letters*, Vol. 15, pp. 1089–1097, 1994.

[10] P. Suetens, P. Fua, and A.J. Hanson. Some computational strategies for object recognition. *Surveys*, Vol. 24, No. 1, pp. 5–62, March 1992.

[11] M J Swain and D H Ballarad. Indexing via color histograms. In *Proc. Image Understanding Workshop*, pp. 623–630, 1990.