Content Based Image Retrieval in Presence of Foreground Disturbances

Rajashekhar and Subhasis Chaudhuri
Department of Electrical Engg.
Indian Institute of Technology-Bombay
Mumbai - 400076. INDIA.
{raja, sc}@ee.iitb.ac.in

Abstract

In this paper we analyze the image retrieval problem in presence of possible foreground disturbances. The foreground may be irrelevant for the retrieval but it occludes the background and hence reduces the retrieval accuracy. We propose the use of a video as a query so that the moving foreground can be extracted. The segmented foreground region is subsequently filled into increase the retrieval accuracy. The performance of a retrieval scheme under foreground disturbance is presented here.

1. Introduction

Use of low level visual features like color, texture, shape, etc, have drawn much attention in the area of content based image retrieval (CBIR). Numerous techniques about efficient image indexing and retrieval from databases have been proposed. Color is often considered as a major feature for indexing because of its role in vision and in identification and discrimination of objects. Color histogram [9] has been shown to be robust to the changes in the object's orientation, scale and viewing position. Authors in [4] proposed the use of color correlogram instead of color histogram. In order to include the local intensity variation, Jhanwar *et al.* presented in [7] a translation and illumination invariant retrieval scheme using motif coocurence matrix. It uses an optimal Peano scan to encode the image. Texture is also an important image attribute which captures the local characteristics of an image. Wavelet based representations have been proposed in [8] for textural feature extraction and its use while doing the texture image retrieval.

One can notice from the CBIR literature that several authors have explored the usefulness of multiple features. In [5] Jain and Vailaya propose to mix color and shape features. The color histogram has been used to index the color feature. Edge histograms and invariant moments have been used for shape representation.

Earlier research on information retrieval is mostly based on using an image as a query for image retrieval and a video for video retrieval. When the user has a video rather than an image as a query, the interest lies in retrieving images with contents similar to the background of the video. This inspires us to come up with a meaningful and an efficient CBIR system by removing possible foreground disturbances, if any, due to moving objects in the video. Often we are interested in retrieving images similar to the background in home video, but the background may not be available completely as there could be some foreground objects in the scene, which are occluding the background. Extracting a frame (could be a key frame) from this video and using it as the query image usually results in a very poor retrieval precision due to the presence of the foreground object. Although a great deal of work has been done during the past many years on CBIR, to our knowledge, researchers have not studied the effect of foreground disturbances in image retrieval. In this paper we study the problem of image retrieval using a video by effectively removing the foreground disturbances. The main contribution of this paper is the investigation of foreground disturbances and how they affect the image retrieval performance. We also contribute towards minimizing the effect of foreground disturbances by filling in the segmented foreground with the pixels of the background scene using the successive frames of the given video. For a given video query we segment the foreground object by applying a foreground subtraction technique. Many algorithms have been proposed in the literature to segment the moving objects in a video sequence. Davis *et al.* presented a non-parametric model called kernel density estimation technique to separate the fore and background objects in [2]. We use the concept of change detection principle to capture the temporal information of the video. The temporal information over a set of video frames is used to extract the foreground object. The foreground subtraction is followed by gradually filling in, if possible the subtracted region using the neighboring pixels of the background scenes from the successive frames. Subsequently we obtain the completely filled in background scene.

(a) (b) (c) (d)

Figure 1. (a) Current frame (I_D) of a query video ($\psi \approx 0.30$), (b) its initial object mask (c) final object mask (d) background image after foreground masking.

In this study we consider the color and texture as the feature for CBIR purposes. Color and texture similarities are integrated with appropriate weights. Experimental results show that elimination of the disturbances due to foreground improves the precision and recall rates significantly with the further enhancement after filling in the segmented foreground region.

2. Foreground subtraction

We exploit the background registration technique as discussed in [1] to segment the moving objects from the query video. Basic idea of this algorithm is change detection. We compute the frame difference mask by thresholding the difference between two consecutive input frames. Based on the history of the frame difference masks (FDM) of several frames, we construct background registration mask (BRM) by considering pixels which are not moving for a long time. Then we compute the background difference mask (BDM) by comparing current input image and the background image stored during the registration step. Using the BDM, FDM, and BRM we construct the initial object mask shown in fig 1(b). Subsequently, the initial object mask is filtered to obtain the final object mask shown in fig 1(c). Because of the filtering operations, the masked out region is typically larger than the actual occluded region. For example, in fig 1(a) the foreground occupies about 30% of the total image size while the masked region is about 34% of the total area (see fig 1(d)).

3. Quantification of disturbance

In order to investigate and quantify the effects of foreground clutters in CBIR, we studied the robustness of the feature matching process during retrieval due to outliers in the feature. The features generated from the occluding region are the outliers. We analyze the sensitivity of the matching process due to presence and exclusion of the occluding region. In addition, we quantify the gain in the matching process when the segmented foreground clutter is filled in using the background sprite. Let us denote an image by I with the subscripts D and BF denoting the image with foreground clutter, and after background filling through sprite generation, respectively. Let f be an operator that works on the image I and generates the corresponding feature vector υ. Thus, υ_D and υ_{BF} are the corresponding features extracted from I_D and I_{BF}, respectively.

Mathematically,

$$\begin{aligned} f &: \mathrm{I}_D \longrightarrow \upsilon_D, \\ f &: \mathrm{I}_{BF} \longrightarrow \upsilon_{BF}, \end{aligned}$$

Let I_I be the ideal query image devoid of any foreground disturbance (see fig 4(d)), and υ_I be the corresponding feature set which would have been ideal for the CBIR applications. Similarly, let I_F be the corresponding foreground image that acts as a disturbance. Please note that we do not restrict ourselves to having chosen a particular feature extractor f. Also, it may be noted that if one is able to do a good background filling, then $\mathrm{I}_{BF} \approx \mathrm{I}_I$.

For a typical retrieval problem we compute the distance of the feature υ_I to those of the images in the database. We now want to quantify what would happen if one has to use either υ_D or υ_{BF} as the feature instead of υ_I. In order to make the problem mathematically tractable, let us assume that the ideal query image and the foreground image causing disturbance are both statistically stationary random processes so that the feature set υ is invariant to the choice of location in the image. Let us assume that $0 \leq \psi \leq 1$ is the fraction of the size of the image I_I corrupted with foreground disturbance.

Thus,

$$\upsilon_D = f(I_D) \quad = \quad \psi\upsilon_F + (1-\psi)\upsilon_I. \tag{1}$$

Hence the similarity measure (assuming an Euclidean distance) with respect to the ideal image I_I is given by

$$\begin{aligned} \mathrm{dist}(\upsilon_D, \upsilon_I) &= \| \psi\upsilon_F + (1-\psi)\upsilon_I - \upsilon_I \|, \\ &= \| \psi\upsilon_F - \psi\upsilon_I \| \; = \; \psi \| \upsilon_F - \upsilon_I \|, \\ &\stackrel{\triangle}{=} \psi \mathrm{M}, \end{aligned} \tag{2}$$

where M is the distance between the foreground and the background. Since the occluding foreground is quite different from the background, M is typically large. Thus, depending on the amount of the background masking ψ, the features of the query image deviates quite drastically from those of the ideal background image.

Let us now see what happens to the similarity measure for an arbitrary image I_{dB} in the database. We may model the image I_{dB} as a mixture of two processes.

$$\mathrm{I}_{\mathrm{dB}} \quad = \quad \beta\mathrm{I}_A + (1-\beta)\mathrm{I}_I, \tag{3}$$

where I_A is a statistical perturbation that makes I_{dB} different from the ideal query image I_I, and $0 \leq \beta \leq 1$

defines the mixing proportion. If β is small then the database image is quite similar to the query image. Assuming feature extractor to be linear operator we obtain $\upsilon_{\mathrm{dB}} = \beta\upsilon_A + (1-\beta)\upsilon_I$. Thus, the similarity measure between the ideal query image and the database image is given by

$$\begin{aligned} \mathrm{dist}(\upsilon_{dB}, \upsilon_I) &= \parallel \beta\upsilon_A + (1-\beta)\upsilon_I - \upsilon_I \parallel, \\ &= \beta \parallel \upsilon_A - \upsilon_I \parallel = \beta \mathrm{M_o}, \end{aligned} \quad (4)$$

where $\mathrm{M_o}$ is the distance between the ideal query image to its similar image in the database and typically $\mathrm{M_o} \ll \mathrm{M}$. If β is small, $\mathrm{I_{dB}}$ is declared to be similar to the query image I_I. If one now, instead, uses the foreground corrupted image I_D as the query, the similarity measure becomes

$$\begin{aligned} \mathrm{dist}(\upsilon_{dB}, \upsilon_D) &= \parallel \beta\upsilon_A + (1-\beta)\upsilon_I - \psi\upsilon_F \\ &\quad -(1-\psi)\upsilon_I \parallel, \\ &= \parallel \beta(\upsilon_A - \upsilon_I) + \psi(\upsilon_I - \upsilon_F) \parallel, \\ &\leq \beta \parallel \upsilon_A - \upsilon_I \parallel + \psi \parallel \upsilon_F - \upsilon_I \parallel, \\ &= \beta\mathrm{M_o} + \psi\mathrm{M}. \end{aligned}$$

Using the fact that both $0 \leq \beta, \psi \leq 1$ and $\mathrm{M_o} \ll \mathrm{M}$ we obtain

$$\mathrm{dist}(\upsilon_{dB}, \upsilon_D) \lesssim \psi\mathrm{M}, \quad (5)$$

Thus we observe that if the foreground disturbance is almost negligible, i.e, $\psi \to 0$ then one is, indeed, able to retrieve similar images from the database. Else the distance $\mathrm{dist}(\upsilon_{dB}, \upsilon_D)$ is quite large even for an image which an user considers to be quite similar to the given query and we end up retrieving irrelevant images. We perform some experimental analysis in quanifying the effect of ψ (the foreground clutter) in the results section. If we are able to detect the foreground clutter, we can mask the region and compute the feature vector only for the background region. Thus, in effect, we try to make $\psi = 0$ in equation 5. As expected, the retrieval accuracy would be enhanced by masking out the clutter region. However, there are two issues that tend to reduce the achievable retrieval accuracy:

(i) Since a part of the image is masked out, the sample size over which the feature υ is computed could be quite small and, the image typically not being homogeneous, the computed feature may be partly different from the true feature vector υ.

(ii) If one is using image texture as the feature, where we make use of the neighborhood properties of individual pixels, computation of the texture near the masking boundary would be erroneous. For ψ being large, this pulls down the accuracy quite significantly. If one is using a point operator such as color histogram as the feature, one does not encounter this problem. However, such features offer a poor retrieval accuracy.

A better option to achieve a higher retrieval accuracy is to perform a filling of the masked region through the generation of background sprite from the query video prior to extracting the features. This process circumvents the problems discussed above and, thus, offers a much better retrieval performance. All these issues are illustrated through experimental results in section 7. It may be mentioned here that depending on the available video, it may not be always possible to fill up the entire masked region.

4. Color as a feature

Color is considered as the most dominant and distiguinshing visual feature. In CBIR, the color histogram is the most commonly used color descriptor. The color histogram describes the global color distribution in an image. The rotation and translation invariant properties of the color histogram motivates us to use the color histogram as one of the featutre vectors in the feature space. As discussed in [5] we extract the global color characteristics of an image by computing three separate 1-D normalized histograms (R,G, and B). As discussed in [5] we use Euclidean distance to compute the color similarity $d_c(I, Q)$ between the query image Q and the database image I.

5. Texture as a feature

Although color is a distinguishing visual attribute, two images with different textures may have identical color histograms. Color attribute may not be able to capture the complete global characterstics of an image. It motivates us to combine the texture with the color attribute. We capture the texture characterstics of an image using the motif cooccurence matrix (MCM) as discussed in [7].

According to this algorithm every database image is divided into 2×2 pixel grids. Each grid is replaced by an optimum scan motif as discussed in [6], which results in the formation of motif transformed image (MTI). We get MTI of size $N/2 \times N/2$, for an image of size $N \times N$. We use this transformed image to get the motif cooccurence matrix which encodes the relationship between intensity variation along specified scan directions in the image. The MCM of the image itself acts as the texture feature vector (see [7] for more details). These MCM feature vectors are subsequently used for computing textural similarity $d_t(I, Q)$ between the query and the database image.

6 Integration of color and texture attributes

Our experimental results show that a single image attribute is not able to give good retrieval rates. In order to improve the performance, we now need to integrate the results due to both color and texture similarities with appropriate weights. We define total similarity measure between query image Q and database image I as,

$$d_{total} = \alpha d_c + (1-\alpha) d_t, \quad (6)$$

Figure 2. Retrieved images in presence of foreground disturbance ($\psi \approx 0.30$) based on both color and texture similarity.

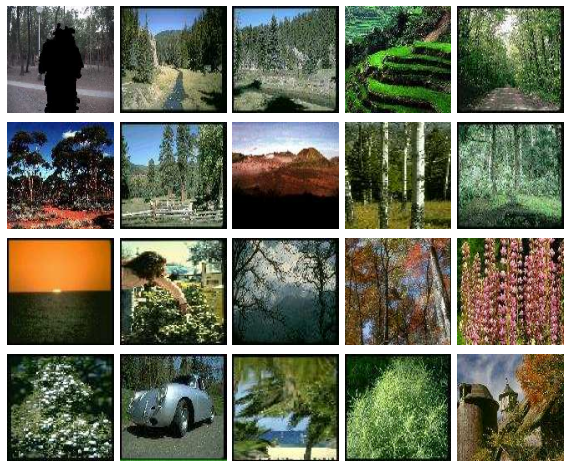


Figure 3. Retrieved images after masking the jogger for a scene with $\psi \approx 0.34$.

where d_c and d_t are the respective color and textural similarities and $0 \leq \alpha \leq 1$ is an appropriate weight based on the relative importance of color and texture features.

7. Effect of disturbance

In order to investigate and quantify the effects of foreground disturbance in CBIR, we conducted experiments on a COREL image database of size 6000. We captured various video clips each one comprising approximately 250 frames. We extracted the foreground object using the background subtraction algorithm discussed in section 2. For example fig 1 (a) shows a video clip where a person jogs away from the camera. The cardinality of the foreground mask changes as the person moves away from the camera. For evaluation purposes one must select a frame which contains the foreground of less cardinality in comparison with the background. If the cardinality is quite large, the feature space will be badly affected by the foreground clutter and the retrieval accuracy will suffer. We plan to study the following in this paper:

(a) (b) (c) (d)

Figure 4. Frames obtained during the process of filling in the segmented region of the image in fig 1(d).

Figure 5. Retrieved images using a query obtained after filling in the segmented region of the image in fig 1(d).

- Effect of foreground occlusion in retrieval accuracy,
- Improvement achieved through the foreground subtraction, and
- Usefulness of foreground filling through video manipulation.

In order to quantify the effects, we use the combined color and texure features as given in equation 6 with $\alpha = 0.45$ for retrieval purposes.

In all experimental results the image displayed first is the query. Ranking begins after the query image and goes from left to right and top to bottom. We now plan to study the effect of the amount of foreground occlusion (ψ) while doing CBIR. In fig 1(a) jogger occupies $\psi = 30\%$ of the natural background scene. Fig 2 shows the retrieved images when this background occluded image is used as a query image. We show the top 20 retrieved images that are similar to the given query image. There are many irrelevant images and even some of the relevant ones rank poorly during the retrieval. It is very clear from the results that the retrieval accuracy is badly affected by the foreground disturbances.

We now study the utility of the foreground removal process while doing the CBIR. We consider the issue of masking out the background scene and its effect on the retrieval

accuracy. Here masking means that the darkened region is not used while computing the color and textural features (see figure 1(d)). Fig 3 shows retrieved images when this is used as a query image. This query is devoid of the foreground clutter, and the size of the background region over which the feature v computed is large when compared to the clutter region. Therefore we obtain a good number of relevant images. Compare this to the results given in fig 2 when the disturbance was not masked out. However, there are two issues that tend to reduce the achievable retrieval accuracy as disucssed in section 3.

In order to further enhance the accuracy we fill in the segmented region using neighboring pixels of the background scenes of the video frames as shown in figures 4(b),(c) and (d). One may use the concept of sprite generation for this purpose [3]. Subsequently we obtain the completely filled background from the successive video frames (see fig 4(d)). After background filling (see query image in fig 5) the background scene is efficiently described by the color and texture features. This further enhances the retrieval efficiency of the CBIR scheme. Experimental results based on background filling are shown in figure 5. Based on these results we justify that retrieval accuracy after background filling is higher than the accuarcy obtained without filling (see figure 3).

We performed the retrieval evaluation in all the cases using the standard evaluation benchmarks such as precision and the recall rates. Precision rate is defined as the fraction of the retrieved images which are relevant. Recall is the fraction of the relevant images which have been retrieved. Figures 6 and 7 display the precision and recall diagrams for with and without the removal of the disturbing region for varying cardinality (size of the foreground). This shows that there is a substantial reduction in both precision and recall rates when the disturbance is not removed from the query video. These plots also substantiate our claim that, if possible, the foreground disturbance must be removed in order to design a meaningful CBIR system. Figure 8(a) describes the behavior of recall rate for the increasing cardinalities of the foreground. The recall curve indicates that the rate decreases when the cardinality of foreground increases. We also notice an improvement in the recall rate after removing the disturbing region. We have shown this for two query examples given earlier. Similarly figure 8(b) displays the average recall rates for increasing amount of filling of the segmented region. This curve shows a significant improvement in the retrieval accuracy compared to the rates obtained due to the background scene with or without the foreground removal.

8. Conclusion

In this paper we presented a CBIR approach using the video as a query to reduce the effect due to any possible foreground disturbance in the scene. The use of video allows us to determine what are the moving object(s) in the scene and how it (they) can be segmented out. Hence the features extracted from the foreground subtracted query image are more meaningful and lead to a much improved accuracy in the CBIR system. In this study we have used a combination of color and texture as the feature set, but one may use any other feature set also. It may be noted that if one is interested in retrieving images that match the foreground and not the background, we can simply swap the definition of the foreground and background. We performed an evaluation of the retrieval system both in presence after removal of the foreground disturbance and also after filling in the segmented foreground region. Performance evaluation is done using standard benchmarks such as precision and recall. We demonstrated the improvement in the retrieval accuracy after foreground segmentation. Subsequently we have also shown a further enhancement in the retrieval efficiency by filling in the segmented foreground region(s).

References

[1] S. Chien, S. Ma, and L. Chen. Effi cient Moving Object Segmentation Algorithm Using Background Registration Technique. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(7):577–586, 2002.

[2] A. Elgammal, D. Harwood, and L. Davis. Non-parametric Model for Background Subtraction. In *Proceedings of 6th European Conference on Computer Vision*, 2000.

[3] N. Grammalidis, D. Beletsiotis, and M. G. Strintzis. Sprite Generation and Coding in Multiview Image Sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(2):302–311, March 2000.

[4] J. Huang, R. Kumar, M. Mitra, W. Zhu, and Zahib. Image Indexing using Color Correlogram. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 762–768, San Juan and Puerto Rico, June 1997.

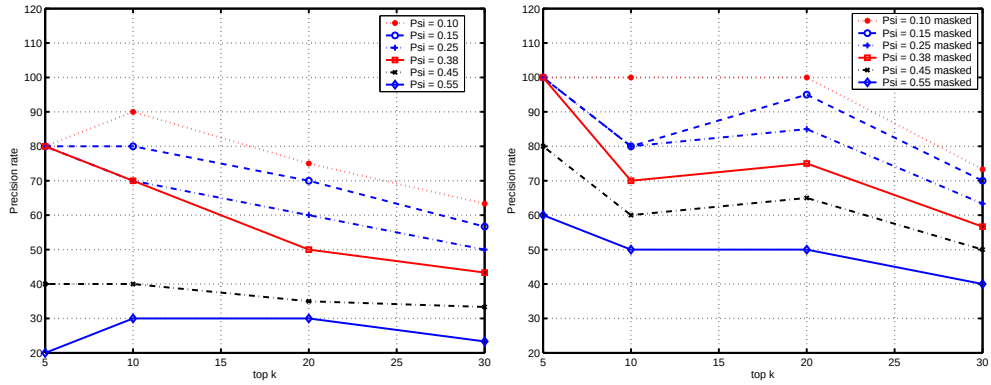
[5] A. K. Jain and A. Vailaya. Image Retrieval using Colour and Shape. In *Proceedings of 2nd Asian Conference on Computer Vision*, pages 529–533, Singapore, 1995.

[6] N. Jhanwar, S. Chaudhuri, G. Seetharaman, and B. Zavidovique. Content Based Image Retrieval Using Optimum Peano Scans. In *Proc. Pattern Recognition*, Quebec, Canada, Agust 2002.

[7] N. Jhanwar, S. Chaudhuri, G. Seetharaman, and B. Zavidovique. Content Based Image Retrieval Using Motif Coocurence Matrix. In *Proc.ICVGIP*, Ahmedabad, India, December 2002 (Enhanced version to appear in IVC 2004).

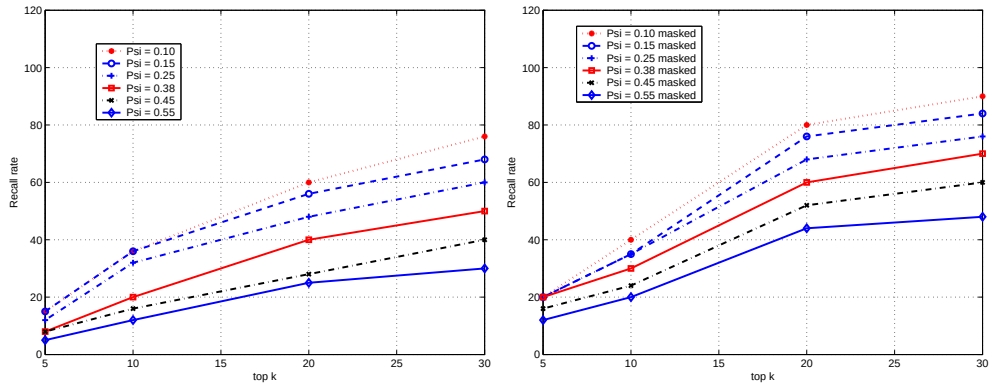
[8] W. Ma and B. Manjunath. A Comparison of Wavelet Transform Features for Texture Image Annotation. In *Proc.IEEE Conference on Image Processing*, Washington D.C, USA, October 1995.

[9] M.J.Swain and D.H.Ballard. Color Indexing. *International Journal of Computer Vision*, 7(1):11–32, September 1991.



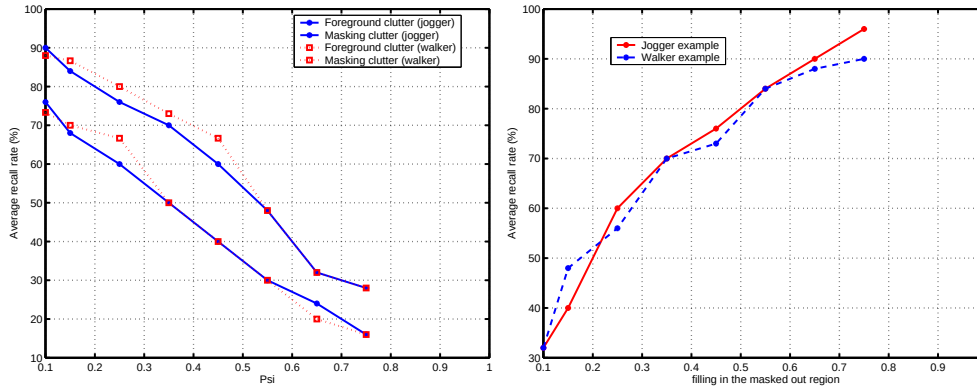
(a) (b)

Figure 6. (a) Precision curves for the CBIR system in presence of varying levels (ψ) of the foreground disturbance, (b) the same curve after the removal of foreground disturbance through masking.



(a) (b)

Figure 7. (a) Recall curves for the CBIR system in presence of the foreground disturbance, (b) the same curve after the removal of foreground disturbance for various cardinalities.



(a) (b)

Figure 8. (a)Average recall rate with the increasing size of the clutter region for two different queries, (b) Average recall rate with the increasing amount of filling of the segmented foreground region.