Video Retrieval Using Color and Spatial Information of Human Appearance

SofinaYakhu and Nikom Suvonvorn

Department of Computer Engineering. Faculty of Engineering, Prince of Songkla University, Songkhla 90112, Thailand

Abstract: The surveillance system manipulates large amounts of video data especially for searching and querying the objects of interest, systematically acts as the content-based video retrieval (CBVR) system. Challengingly, the video retrieval technique must be accurate, reliable, and not complex due to time consuming according to amounts of data. In this paper, we propose a novel method of object-based video retrieval using color and spatial information of human appearance, applied for searching human object in the video surveillance system. Experimentally, our method archives at high precision and recall rate.

Key words: Video retrieval, Spatial information, Color Histogram, Surveillance System

1. Introduction

In surveillance system, the amount of video content has grown extensively during days due to video recording continuously for long periods of time. Accordingly, a video retrieval system is needed systematically for such application, for example, searching or querying the specific object from video. Many researches have investigated various approaches. Jeong [1] used global color histogram indexing method for image retrieval providing reasonable retrieved images. Color histogram can be both an advantage and a disadvantage is their lack of spatial information. To solve this problem, Huang et al. [2] investigate color correlogram that integrates both color information and space information which more stable to color change but more complex. Some research focus on video retrieval, Liang-Hua et al. [3] proposed a video retrieval algorithm based on the integration the color and motion features. They analyze all frames within a shot to construct a compact representation of video shot. This approach able to fully exploit the spatio-temporal information contained in video. Besides, another approach is proposed by Grauman et al. [4] which maps unordered feature sets to multi-resolution histograms and computes a weighted histogram intersection in feature space. For the matching, they used pyramid matching that is believed as one of the best matching algorithms for image retrieval and recognition. Pickering et al. [5] present global features were extracted from key frames and used as the basis for content-based retrieval using a vector space model, k-nearest neighbors and boosting.

In our research, we consider human in video sequence as interested objects, which are extracted from frames of the shot. We emphasize on object matching by comparing between query object and objects in shot. We propose an object-based video retrieval technique applied to video surveillance system for querying human image from large video database which considering both color and spatial information.

2. Proposed System

The remaining sections of paper will discuss on the proposed method of object based retrieval technique applying for surveillance applications. Accordingly, the retrieval on human physical appearance as moving

object in the video surveillance is emphasized. Following the application context, images of suspected human found in the observing area will be used as image object for retrieving the similar objects from a huge database of video. The overall of proposed system is shown in Fig 1, which is divided into three main parts: moving object detection and two-stage decision process using object similarity measurement. The two-stage decision of object similarity is established in cascade in order to progressively reject the non-query objects.

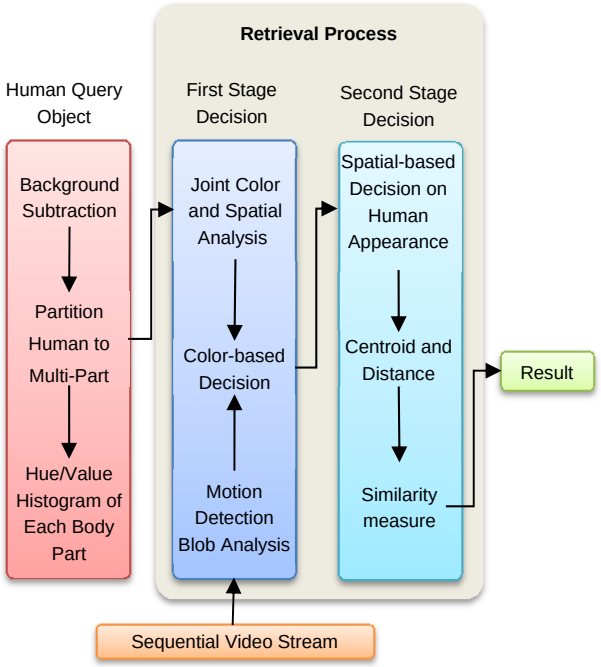


Fig.1 Overall System

3.1 Query Object: Human

In this stage, query object are objectively chosen by user, obtaining only full human body with some background as shown in the Fig.2. The basic idea of querying object is to generate a candidate partition for each human body for searching. Two types of human partitions are considered using the anatomical principle of human body [6], in Fig.2(left) human body can be identified globally into three components: upper body (13%), middle body (40.3%) and lower body (46.7%), which are correspond to the head, torso and limb respectively. Furthermore, in fig.2 (right) human body can be separate into four components: upper body (13%), middle body (37.3%) first lower body (18.7%) and second lower body (35%), which are correspond to the head, torso, upper limb and lower limb respectively.

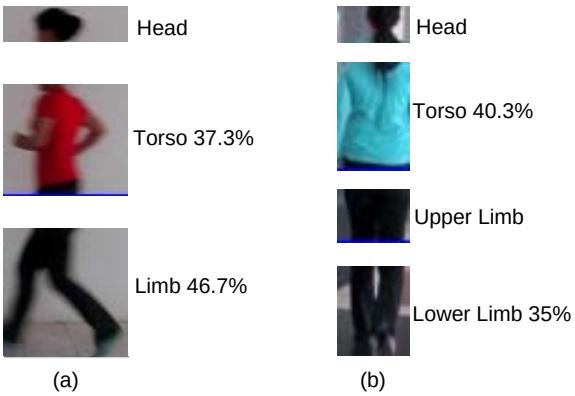


Fig.2 Separating human body to (a) three components and (b) four components.

The HSV color space is applied in order to identify chimerical part of colors of the observing human – the brightness component is ignored. For the specific query object, a set of color histogram $H^Q_{body} = \{H^Q_i | i = 1,2,\dots,n\}$ of the human body is determined where H^Q_i denotes the histogram of each part, such as, histogram of head H^Q_1, torso H^Q_2 and limb H^Q_3. Note that black color or dark gray color are not represented in Hue component, which can appear often in human's clothes. We then use the combination of histograms from the hue and value component of HSV color space by the following definition:

$$H^Q_i = \begin{cases} H^Q_i(h) & if\ V_{max} = arg\max_v H(v) > th \\ H^Q_i(v) & otherwise \end{cases} \quad (1)$$

where $H^Q_i(h)$ and $H^Q_i(v)$ denotes the Hue and Value histogram respectively. Objectively, we use the Hue histogram for representating body parts when the maximal value of Value histogram is higher than a theshold *th*. Notice that when value of Value histogram is small, the image became dark.

3.2 First Stage Decision: Color Comparison

In the first stage of decision, we try to determine which moving regions are representing by the same color of queries object. Two steps are required: motion detection and histogram projection. In the motion

detection step, we use a simple technique, named frame difference, to detect a movement object of the video frame as following equation.

$$|frame_i - frame_{i-1}| > threshold \quad (2)$$

The foreground pixels are then segmented into regions using connected component as region of movement R^k , where k = 1,2,..N. However, the movement region may be not cover human body. We used blob detection [7][8] to selected boundaries movement. The movement region is defined as R_m^k.

Then, the back-projection technique [9] is applied in order to extract the image regions using the histograms $H_i^Q$of head, torso, and limb. Note that the histogram of each frame in the shot is defined H_m^k. To get the back-projection image, we compute the histogram ratio R_j^k where j is histogram bin as shown in equation (3).

$$R_j^k := \min\left(\frac{H_i^Q}{H_m^k}, 1\right) \quad (3)$$

Then, values R_j^k is replaced to image values which values to perform backprojected image *b*. The backprojected image is convolved by a mask, which is compact objects of unknown orientation could be a circle with the same area as the expected area subtended by the object. The peak in the convolved image is the expected region R_b^k. To accept or reject regions, we verify very frame in a shot that the following condition must be satisfied.

$$R_h^k = \{ R_b^k \cap R_m^k \,|k = 1 \ldots \mathrm{N}\} \quad (4)$$

Notice that R_h^k is acceptance regions, which is defined by the moving regions that have the same color properties of head or torso or limb. For the regions that this condition is not verified will be rejected.

3.3 Second Stage Decision: Spatial Comparison

In this stage, the regions obtained from the last decision will be verified spatially using information of human body. To achieve the objective, we determine the centroid of each region. The baricenter is computed from the ratio of pixels having color *c* in image *I*. Note that $\Lambda_c^I := \{(x,y) \in I : I[x,y] = c\}$ [10] defined as the set of pixels of *I* having the same color *c*. The centroid of human region R_h^k $(\bar{x}_{R_h^k}, \bar{y}_{R_h^k})$ is shown in equation (5) and (6).

$$\bar{x}_{R_h^k} = \frac{1}{n}\frac{1}{|\Lambda_c^I|}\sum_{(x,y)\in A_c} x \quad (5)$$

$$\bar{y}_{R_h^k} = \frac{1}{m}\frac{1}{|\Lambda_c^I|}\sum_{(x,y)\in A_c} y \quad (6)$$

For matching between human query and each frame in shot, we verified in relations between the three human component that is three location of centroid. Therefore, the distance between centroid was significantly. We used euclidean distance technique to compute distance between centroid by the following equation.

$$d\left(R_h^k, R_h^{k+1}\right) = \sqrt{(\bar{x}_{R_h^k} - \bar{x}_{R_h^{k+1}})^2 + (\bar{y}_{R_h^k} - \bar{y}_{R_h^{k+1}})^2} \quad (7)$$

Consequently, any possible of three regions are grouped into human structure. However, the following two conditions must be verified in order to accept as the possible structure: (1) three histograms of grouping regions must include all parts of backprojected image of human body having the same order; (2) the spatial structure of human must satisfy the following condition – the proportion of human structure projected in x and y axis must be quasi the same using torso as a reference point.

$$f(R) = \sqrt{\left(\frac{d\left(R_i^Q, R_{i+1}^Q\right)}{d\left(R_{i+1}^Q, R_{i+2}^Q\right)} - \frac{d\left(R_h^k, R_h^{k+1}\right)}{d\left(R_h^{k+1}, R_h^{k+2}\right)}\right)^2} \quad (8)$$

where $f(R)$ is defined as cost function of object retrieval that compares between ratio of distances of centroid's human query and centroid's extracted image regions. Objectively, when objects are similar, the funtion $f(R)$ give value close to zero, the output will be retrieved. If this second condition is not verified, then the composition will be rejected.

Fig.3 Some of the query objects in different

3. Experimental Results

4.1 Experimental setup

In our experiments, we setup the datasets of human activities in order to evalute performance of our method. Actions of activity with different clothes are demonstrated, such as, walking and running of human showing front view (24 shots) back view (25 shots) and side views (25 shots). Experimentation is done in indoor environment, which is normal case of video surveillance system. Some examples are shown in Fig.3.

We conducted three experiment tasks. In the first retrieval testing, querying human is segmented into three components corresponding to the head (13%), torso (40.3%), and limb (46.7%), (Model I). The second retrieval test, querying human is separated into four components: upper body (13%), middle body (40.3%), first lower body (18.7%) and second lower body (35%), but the lowest body is not considered (Model II). This task will showed performance of result when human wear shorts or skirt. We will compared the performance between these two experimentations. The third testing, retrieving human is performed on the shots including many people considering as a real situation.

4.2 Evaluation method

We evaluate our method by measuring the precision and recall. The pricision is the percentage of true detection with respect to the overall declared event, shown in equations (9). The recall is the true positive rate from detectionand, detailed in equations (10).

$$\text{Precision} = \frac{Number\ of\ relevant\ objects\ retrieved}{Total\ number\ of\ objects\ retrieved} \quad (9)$$

$$\text{Recall} = \frac{Number\ of\ relevant\ objects\ retrieved}{Total\ number\ of\ relvant\ objects\ in\ database} \quad (10)$$

The retrieval results using our method of first task are given in Table 1, showing the number of correct, missed and false detection according to the different datasets. To evaluate the peformance of our method, precision and recall are computed for each testing.

4.3 Results and Discussion

In the first testing with Model I, our method globally provides satisfied result in term of precision: the maximum/ minimum precision are at 1.0 and 0.71 respectively. Accordingly, the recall is also obtained at high rate from 0.68 to 0.96. Additionally, correction rate is high at 92.4% while the false detection rate is very small at 11.8% in the worst case. However, missed detection rate is quite high at 8.12% due to the imperfect result of motion detection and the environment where colors of clothes and background are quasi the same. In particular case, by comparing the performance of the query behaviors, querying the side view of human gives better results than other any testing configurations. Nevertheless, these results are quite logic because the side view may represent the extensive details of front view and back view.

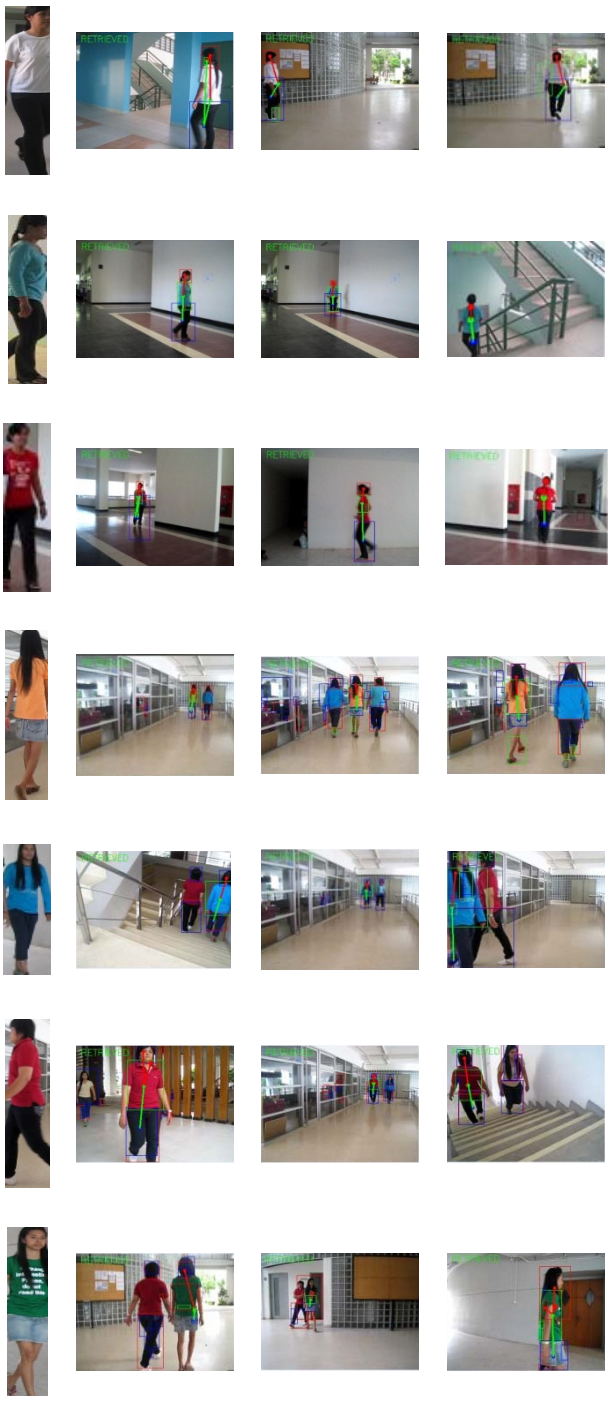
The second testing with Model II when querying human is partition into four parts is shown in Table 2. We can clearly notice that the method give high

values for both recall (ranged from 0.69 to 0.90) and precision (ranged from 0.84% to 0.99%). Globally, the recall is smaller than precision values. Again, we found that color of shirt, skirt, or trouser can effect directly to the results due to the similarity of color to background, for example, white shirt and white background gives recall only 0.41. In the comparaison with Model I, we can notice that in average both recall and precision of the Model II give better results than the Model I, especilly when people wears skirt. This can be explained by the fact that in the first experimentation colors of leg and background become majority of the lower part of body that effect to wrong result.

In the third experimentation, we test how our method can retrieved the specific human from interacting peoples in shots, walking or running. In this case, we applied the Model II of human in the test due to the best performance shown in the last testing. Fig.4 (b) depicts some results, and Fig.4 (a) show querying human object. Table 3 describes the results as values of recall and precision rate in average. So, the maximum and minimum of precision rate are 0.97 and 0.81 respectively, and those of recall are 0.96 and 0.78 correspondingly. Again, we can notice that our method can globally perform a good result. However, some factors can fail our method when people wear clothes with similar color but not the correct one.

6. Conclusion

In this paper, we introduced a novel technique for human object retrieval applying for video surveillance system using color and spatial relation of human physical appearance. The experimentation result shows that our method can retrieve different human actions at high correction rate 92.4%. The query of side view gives the best result at 1.0 of precision rate. Seperating four part of querying human and consider only three upper parts retrieved better when human wear shorts or skirt because of the lowest body have color not interest.



(a) (b)

Fig.4 Query example using Model II of human appearance: (a) quering human and (b) retrieval result

References

[1] Sangoh Jeong. Histogram-based color image retrieval. Psych221/EE362 Project Report, 2001.

[2] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. “Image Indexing Using Color Correlograms”. Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition .Washington, DC, USA: IEEE Computer Society, 1997, PP:762

[3] Kuo-Hao and Liao, Hong-Yuan Liang-Hua and Chin. “An Integrated Approach to Video Retrieval”. Nineteenth Australasian Database Conference (ADC 2008), 2008, PP:49-55.

[4] Kristen Grauman, and Trevor Darrell. “The pyramid match kernel: Discriminative classification with sets of image features”. In Proceedings of the IEEE International Conference on Computer Vision. Beijing, China, 2005.

[5] Marcus J. Pickering, Stefan M. Rüger, David Sinclair, “Video Retrieval by Feature Learning in Key Frames”. CIVR 2002, 2002, PP: 309-317

[6] G. Gaughran W.Dempster. “Properties of body segments based on size and weight”. American Journal of Anatomy,1967, PP:33-54.

[7] A. Ming and H. Ma, "A blob detector in color images". Proceedings of the 6th ACM international conference on Image and video retrieval, Ney York, USA, 2007, PP: 364-370.

[8] J. Matas, O. Chum, M. Urban and T. Pajdla, "Robust wide baseline stereo from maximally stable extremum regions". British Machine Vision Conference. 2002, PP:384–393.

[9] M.J. Swain and D.H. Ballard, “Color indexing”, International Journal of Computer Vision, 1991, pp.11-32.

[10] L. Cinque, S. Levialdi, A. Pellican, and K. A. Olsen. 1999. “Color-Based Image Retrieval Using Spatial-Chromatic Histograms”. In Proceedings of the IEEE International Conference on Multimedia Computing and Systems - Volume 2 (ICMCS '99), Vol. 2. IEEE Computer Society, Washington, DC, USA, 969.

Table 1 Results of Model I.

Query behavior (*Partition 3-component*)	**Number of frames**		**Precision**	**Recall**
	Correct detection	Missed Detection		
- Red shirt-Black trousers				
Front	193	23	0.89	0.99
Back	464	50	0.90	0.99
Side	178	8	0.97	0.99
- Blue shirt-Black trousers				
Front	276	11	0.96	1
Back	334	24	0.93	0.99
Side	293	11	0.96	1
- Green shirt-Black trousers				
Front	336	70	0.83	1
Back	300	68	0.82	0.95
Side	151	15	0.92	1
- White shirt-Black trousers				
Front	171	0	1	0.97
Back	213	39	0.85	0.97
Side	86	19	0.82	0.96
- Yellow shirt-Light blue skirt				
Front	102	21	0.83	0.99
Back	100	23	0.81	0.95
Side	163	18	0.90	1
- Pink shirt-Blue trousers				
Front	178	62	0.74	0.94
Back	532	88	0.86	0.98

Query behavior (*Partition 3-component*)	Number of frames		Precision	Recall
	Correct detection	Missed Detection		
Side	119	14	0.89	0.99
- Pink shirt-Khaki trousers				
Front	185	41	0.82	0.92
Back	287	98	0.75	0.97
Side	175	38	0.82	0.98
- Black shirt-Red shorts				
Front	64	15	0.81	0.82
Back	187	46	0.80	0.99
Side	250	26	0.90	0.99

Table 2 Performance comparison using Model I and Model II.

Costume	Partition Human to Three Parts						Partition Human to Four Parts (consider only 3 upper component)					
	Front view		Back view		Side view		Front view		Back view		Side view	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Red shirt-Black trousers	0.83	0.96	0.83	0.94	0.82	0.89	0.82	0.94	0.84	0.95	0.83	0.94
Green shirt-Black trousers	0.91	0.98	0.89	0.99	0.91	0.99	0.88	0.99	0.88	0.98	0.87	0.99
Blue shirt-Black trousers	0.86	0.94	0.68	0.93	0.88	0.94	0.88	0.91	0.69	0.95	0.89	0.93
White shirt-Black trousers	0.89	0.91	0.90	0.89	0.89	0.89	0.87	0.94	0.41	0.89	0.65	0.92
Emerald shirt-Black trousers	0.88	0.96	0.92	0.97	0.91	0.98	0.86	0.87	0.90	0.97	0.90	0.94
Green shirt- Light blue skirt	0.86	0.90	0.96	0.79	0.85	0.90	0.82	0.96	0.78	0.84	0.89	0.95
Pink shirt-Blue trousers	0.76	0.93	0.77	0.91	0.76	0.94	0.83	0.87	0.80	0.91	0.81	0.98
Pink shirt-Khaki trousers	0.84	0.94	0.78	0.94	0.85	0.96	0.83	0.92	0.86	0.99	0.86	0.98
Black shirt-Red shorts	0.81	0.90	0.79	0.88	0.81	0.88	0.85	0.93	0.81	0.90	0.87	0.98

Table 3 Results of querying human from complex scene.

Many People in shots	Model II (*Partition 4-component and consider only 3 upper component*)		
	Front view	**Back view**	**Side view**
• Dark red shirt/Black trousers			
Average Recall	0.79	0.80	0.88
Average Precision	0.96	0.97	0.91
• Orange Shirts/Light blue skirt			
Average Recall	0. 78	0. 88	0.92
Average Precision	0. 89	0. 86	0.83
• Green shirt/ Light blue skirt			
Average Recall	0. 90	0. 90	0.80
Average Precision	0. 89	0. 81	0.92
• Blue shirt/Dark blue trousers			
Average Recall	0. 87	0. 90	0.96
Average Precision	0. 89	0. 87	0.94

Sofina Yakhu was born in Yala, Thailand, on June 24, 1987. She received the B.Eng (Computer Engineering) in 2008, from Prince of Songkla University (PSU), Hat Yai, Songkla, Thailand. She is currently pursuing M.Eng (Computer Engineering) at PSU in computer vision. Her research interests are in the areas of image processing and computer vision.

Nikom Suvonvorn was born in Trang, Thailand, on November 26, 1976. In 2006, He received a PhD in computer science from l'Université de Paris Sud (XI), Orsay, France. In 2003, He obtained a DEA (Diplôme d'Etudes Approfondies), on Electronic System and Information Processing (SETI) from l'Institute d'Electronique Fondamontale (IEF) at the same university. In that year, He also got another master's degree on computer engineering from Ecole Supérieure de Mécanique et d'Electricité (ESME)-Sudria engineering school, Paris. He is currently a lecturer and research scientist at Department of Computer Engineering (CoE), Faculty of Engineering (ENG), Prince of Songkla University (PSU), Hatyai, THAILAND. His research corresponds to computer vision, image processing, and its related applications. The actual research is emphasized on the OpenVSS project, the next generation multimedia technologies applied for the Surveillance & Smart Environment System