Using Color Histograms to Recognize People in Real Time Visual Surveillance

DANIEL WOJTASZEK, ROBERT LAGANIERE
S.I.T.E. University of Ottawa,
Ottawa, Ontario
CANADA
danielw@site.uottawa.ca, laganier@site.uottawa.ca

Abstract: - This paper presents a surveillance system that detects and recognizes people in indoor scenes. To distinguish between different people, color histograms on a perceptually uniform color space are used. The goal here is to associate a sequence showing a person leaving a room with the previously recorded sequence showing that same person entering.

Keywords: - Recognition Color Histogram Surveillance People Tracking

1 Introduction

In many surveillance applications, recognizing people in the scene is desired. The most unique visual features one can extract from an image of a person are facial features. People recognition, therefore, usually proceeds by extracting some facial features and finding a match from among the features of faces already stored in memory. A limitation to this method is if someone is in the scene but his face is not visible from the camera(s).

Before a system can recognize a person, it must detect and locate them in the image. A common method of person detection in a sequence of images is comparing the current image with a reference or background image. Any pixels in the current image that differ according to some criteria from the background image are labeled as foreground pixels. The foreground pixels are then analyzed and the location of a person is determined using some criteria such as shapes of clusters of foreground pixels. A few approaches to background modeling, foreground segmentation and person location are presented in [1], [2] and [3]. Another method that uses a pattern recognition technique was proposed in [4]. This method involves creating a hidden Markov model of some features extracted from sample images of a person before the system is brought on line. When the system is on line, the Viterbi algorithm is used to segment the image into foreground and background regions.

Once a person is located in an image the next step is to determine who this person is if visual features of this person have previously been acquired. There are two steps to recognizing a person: extracting the proper features and comparing these features to determine if they come from the same person or not. Orwell, et al. [5] combine the foreground pixels which are identified as part of a single person into a vector which is then used as the feature. They then determine if the features extracted from different images of people match using statistical methods such as χ^2 probability function.

The environment that a surveillance system is designed to monitor greatly influences the methods used. For the purposes of visual surveillance the most important environmental factors are the lighting and the expected activity in the scene being observed. One situation where visual surveillance is useful is when we would like to monitor a passage inside a building. This passage could be a doorway, a hallway or any other area that people use to go from one area to another. The application presented in this paper monitors the doorway to our laboratory using a single color camera. More specifically, we want to record any activity in this scene, to keep count of how many people are in the laboratory at any given time and to keep a record of the activity that occurs in the laboratory during a certain period of time. The scene under observation shows the entrance portion of the lab and also includes a view of a door giving access to an adjacent room.

2 People Detection

The first step in people detection is to form the silhouettes of foreground objects in the image. The foreground regions of an image are extracted by a pixel wise comparison between this image and a background image. Any pixels that differ by more than a threshold are labeled as foreground (white) pixels in a binary image. All other pixels are labeled as background (black). Each pixel of the background image is modeled using a Gaussian distribution. The threshold for each pixel is three times the standard deviation for that pixel. Indeed observing an indoor scene with good lighting conditions and with moving objects (people) of relatively important size does not necessitate the use of a more complex background model such as the model presented in [6]. Median filtering is used next to remove any small groups of foreground pixels that most likely are not a part of any object of interest. Finally, closing morphological operations are used to improve the shape of each silhouette. Figure 1 shows the silhouettes of two people formed using this method.

Once the silhouettes have been formed they must be analyzed to see if any of them fit the shape of a person. To do this, local vertical peaks on the boundary of each silhouette are located using Quasi-Topological Codes in a similar fashion to what was done in [3]. From each local peak found, the silhouette boundary is scanned in the left and right directions recording the curvature using a square four element window.

p1	p2
p4	p3

$c = p1 + 2p2 + 4p3 + 8p4$
$p1; p2; p3; p4 = 0 \text{ or } 1$

The value of c determines what the curvature of the boundary at a certain point is. If tracing the silhouette boundary starting from the maximum point in the left and right directions yields a convex curve (downward curve) and then a vertical drop, then this boundary is likely to be the boundary of a head. Next, the width of the putative head is determined by recording the horizontal coordinate of the pixel on the boundary of the head which is furthest left of the local maximum point and the pixel on the boundary of the head which is furthest right of

(a) **(b)**

Fig. 1 Example of silhouette formation. (a) A sample image. (b) The silhouette formed using background comparison.

the local maximum point. The final criteria for determining if each putative head is in fact a person's head is that it must have a body under it. To determine this, a region is defined for each putative head which is bounded horizontally by the extreme horizontal coordinates of the boundary of the head found above, on top by the vertical coordinate of the maximum point and on the bottom by the vertical coordinate of the top plus the width of the head multiplied by some constant.

$b = m_y + k * w$

Where m_y is the vertical coordinate of the maximum point, k is the expected ratio of a person's height to the width of his head, and w is the width of the head.

In this application we limit the height to the length from the top of the head to the beginning of the legs. The possible head is considered above a body if this defined region of the binary image has a high enough percentage of pixels labeled as foreground. Figure 2 shows two examples of this region indicated by the black rectangles in the image.

Fig. 2 A sample image showing two detected persons and the regions of interest.

3 Feature Extraction

In order to monitor the activities that occur in the room under observation, we have to be able to recognize a person that passes in front of the camera. More specifically our goal is to associate a sequence showing a person leaving the room with the previously recorded sequence showing that same person entering the room. To distinguish between different people we decided to use color information. This will work well if the observed people wear a good variety of clothing and if we expect that they will not change their clothing once inside the room. Also since this system monitors an area where people are passing through and therefore are seen from different angles, we choose not to use skin or hair colors. These attributes are more likely to be similar between different people and are more difficult to reliably extract from different angles. Moreover clothing color information is generally radially invariant.

To extract color information, a three dimensional histogram is used. We choose to use a histogram because the histograms of the colors extracted from people wearing different colored clothing are different whereas histograms of the same person taken at different times are more similar to each other. Two dimensions describe the chrominance, $q_{u'}$ and $q_{v'}$, defined by the CIE Uniform Chromaticity Scale and the third dimension describes the luminance defined by the y component of the CIE XYZ color space. Note that only pixels which are labeled as foreground and do not represent a part of the person's head will be added to the histogram.

$$y = 0.2127R + 0.7152G + 0.0722B$$
$$x = 0.4125R + 0.3576G + 0.1804B$$
$$z = 0.0193R + 0.1192G + 0.9502B$$

$$q_{u'} = \frac{4x}{x+15y+3z}$$

$$q_{v'} = \frac{9y}{x+15y+3z}$$

Where R, G, and B are tristimulus values from the Rec 709 Primaries.

The chrominance defined above are perceptually uniform which means that two colors which are at a fixed Euclidean distance from each other on the ($q_{u'}$, $q_{v'}$) plane will have the same relative perceptual difference to a human no matter where these colors are located on the ($q_{u'}$, $q_{v'}$) plane. This characteristic allows one to decide numerically which pair of colors look more alike given several colors.

The domain of each dimension of the histogram is chosen to be the range of the corresponding color component. The number of bins and the bin boundaries are chosen off line and are fixed. These parameters are fixed to make accumulating color information from several images of the same person in the image sequence into a single histogram more efficient. The number of bins, in other words the coarseness of quantization, for each dimension is chosen depending on how much computational power is available. The higher the number of bins, the better the recognition results and the more computational power required. Of course when the number of bins is fairly large, increasing this number by any significant amount will not significantly improve the recognition results. The bin boundaries are chosen to be uniformly spread across the domain of each dimension.

Another important factor in extracting the color of a person's clothing is what region in the image to extract the color from. This region should most likely be void of any skin or hair. A simple region which satisfies this criterion is a rectangular region below the person's neck. An example of this region is indicated by the black squares in figure 2 excluding the area containing the head and neck of the person. Choosing to extract the color from this region greatly reduces the possibility of colors from other people in the image being mixed with the colors extracted from the person of interest due to overlap of body parts. The area of the head and neck of a person is approximated by a rectangular region of width equal to that of the head and height proportional to the width.

When a person passes through the scene they are tracked and the color information discussed above is accumulated from every image of the person as they pass through the scene into the histogram. The histogram is then normalized so that each bin now contains the percentage of the total number of pixels accumulated in the histogram. Tracking is accomplished by using color histograms to compare a person detected in an image to any person detected in recent previous images to determine correspondences. So if the same person is detected in

several consecutive or nearly consecutive images, the information about this person extracted from these images is labelled as belonging only to this person. Figure 3 shows the path of the head of a person who is leaving the lab as was determined by tracking this person.

Fig. 3 An image showing the path taken by a person.

4 People Matching

To compare two histograms a measure of dissimilarity is computed using the Earth Mover's Distance, EMD, presented in [7]. For this measure, one of the histograms, called the source histogram, is like several piles of earth on a field where each pile represents a bin in the histogram, the mass of earth in each pile represents the value of the histogram at the corresponding bin and the field represents the domain of each dimension of the histogram. The other histogram, called the destination histogram is like several holes in a field where a hole represents a bin in the histogram, the volume of each hole represents the value of the histogram at the corresponding bin and the field is the same as the field described for source histogram. The EMD is the minimum energy required to fill the holes in the field with the earth from the piles in the field.

When several sequences of people passing through the scene are compared to a new image sequence in this way, the two sequences which yield the lowest value of the EMD are considered the most likely to be the same person.

5 Results

To measure the effectiveness of this matching technique we proceeded as follows:

1. Image sequences of twenty different people are captured. Two sequences are captured for each person: one of the person entering the lab and one of the person leaving the lab.

2. n out of the twenty people are randomly selected and these people are assumed to be in the lab.

3. The matching technique is used to compare a person's leaving sequence to each entering sequence of the people in the lab before this person left. If the minimum value of the EMD is generated from comparing two sequences of the same person then recognition is successful. Else the recognition is unsuccessful.

4. Step 3 is repeated using a different person's leaving sequence until each of the n people has been the leaving person.

5. Steps 2 to 4 are repeated ensuring that a different set of n people are chosen each time until each of the twenty people have been included in at least one of these sets.

6. Steps 2 to 5 are repeated for each value of n ranging from two to twenty.

Recognition rate for each value of n, R_n , is then calculated.

$$R_n = N_n^s / N_n^t$$

N_n^s**:** The number times step 3 is performed for the same value of n and results in a successful recognition.

N_n^t**:** The total number of times step 3 is performed for the same value of n and results in either a successful recognition or not.

Figure 4 shows a graph of recognition rate, Rn versus the number of people in the lab, n.

6 Conclusion

A method to detect and recognize people in indoor scenes has been presented. To detect people, the use of silhouette shape cues was used. To recognize people, color histograms were used on a perceptually uniform color space. These methods were successful

at least eighty percent of the time for recognizing people who reenter the scene.

References:

[1] M. Lee, Detecting People in Cluttered Indoor Scenes, *Proceedings of CVP*R, 2000, pp. 804-809.

[2] I. Haritaoglu, D. Harwood, L. Davis, Hydra: Multiple People Detection and Tracking Using Silhouettes, *Proceedings of Second IEEE International Workshop on Visual Surveillanc*e, 1999, pp 6-13.

[3] J. Heikkila, O. Silven, A Real-Time System for Monitoring of Cyclists and Pedestrians, *Proceedings of Second IEEE International Workshop on Visual Surveillanc*e, 1999, pp 74-81.

[4] G. Rigoll, B. Winterstein, S. Muller, Robust Person Tracking in Real Scenarios with Non-Stationary Background Using a Statistical Computer Vision Approach, *Proceedings of Second IEEE International Workshop on Visual Surveillanc*e, 1999, pp 41-47.

[5] J. Orwell, P. Remagnino, G.A. Jones, Multicamera Colour Tracking, *Proceedings of Second IEEE International Workshop on Visual Surveillanc*e, 1999, pp 14-21.

[6] M. Harville, G. Gordon, J. Woodfill, Foreground Segmentation Using Adaptive Mixture Models in Color and Depth, *Proceedings of IEEE Workshop on Detection of Events In Vide*o, 2001, pp 3-11.

[7] Y. Rubner, C. Tomasi, L. J. Guibas, The Earth Mover's Distance as a Metric for Image Retrieval. *Technical Report STAN-CS-TN-98-86, Departement of Computer Science, Stanford Universit*y, 1998.

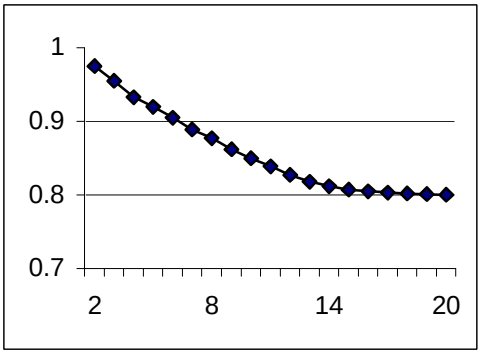

Fig. 4 Graph showing recognition rate vs. number of people in the sample set.