Tracking and Recognizing People in Colour using the Earth Mover's Distance

DANIEL WOJTASZEK, ROBERT LAGANIÈRE
S.I.T.E. University of Ottawa,
Ottawa, Ontario, Canada K1N 6N5
danielw@site.uottawa.ca, laganier@site.uottawa.ca

Abstract

This paper proposes a method that tracks, and recognizes people in indoor scenes. Background subtraction and foreground silhouette analysis are used to detect people. The colour of each person's clothes is used as a distinguishing feature for the purpose of tracking and recognizing people. Clothing colour is extracted using histograms on a luminance and perceptually uniform chrominance colour space. A measure of dissimilarity between different histograms is computed using the earth mover's distance. This approach has been used in an application consisting in monitoring people entering or leaving a room.

1. Introduction

A common task for computer vision systems is monitoring human activities in a scene. Many methods have been developed to fulfill this objective in the context of visual surveillance, telepresence and others. The basic steps required to accomplish this task are the followings: image acquisition, background extraction, moving objects segmentation, person detection and person tracking. The information obtained from the execution of these steps can then be analyzed to identify the activities that take place and/or to attempt to recognize the detected people.

The goal of this paper is to propose a method to perform tracking and recognition of moving people using color histograms. While a person is tracked, characterizing information is recorded such that the next time this person appears in the scene, he or she will be recognized by the system.

But before a system can track and recognize persons, it must detect and locate them in a sequence of images. The first step is therefore to model the background of the observed scene such that when a new element appears in the scene it will be detected and extracted. A frequently used strategy consists in modeling the temporal observations of the background at each pixel location as a Gaussian mixture [1]. This model is dynamically learnt and updated as the scene is observed. Any pixel value that cannot be explained by this model are then classified as belonging to foreground objects.

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. For example, [2] used the fact that that the top of each shoulder and head have relatively little curvature in the vertical direction when compared to the curvature of the sides of the head. The shape of a head and shoulders can then be detected by locating regions with a low value of the horizontal derivative delimitated by high derivative values. In [3], it is observed that when persons are walking, their head are usually almost directly above their torso. To detect people walking through the scene the extreme convex points of curvature and the vertical projection histogram of each foreground region are extracted. If a peak in the histogram is close to an extreme convex point of curvature then there is a head of a person in this region. [4] uses the fact that the silhouettes of most people who are standing have very similar aspect ratios and areas, and that a person's head is usually almost directly above their torso. Another method which uses a pattern recognition technique was proposed in [5]. 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.

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. In [6], people are recognized by accumulating the color of all the foreground pixels which are identified as part of a single person into a measurement vector. They then determine if the features extracted from different images of people match using a χ^2 probability function. The color representation used is a non-uniformly quantized version of the (y,u,v) color space. Color ratio gra-

dients, that are invariant to object pose and lighting conditions, are used by [7] in a quadtree-based split and merge segmentation to segment an image by texture. A measure if similarity between two histograms is proposed by [8]. It proceeds by determining the intersection between the histograms; the intersection is the sum of the minimum value of each corresponding bin taken over all bins in the histogram. In [9], RGB color moments of varying order and degree are proposed as features to recognize color patterns even with changes in illumination and viewing position.

Section 2 of this paper describes the process of person detection. Section 3 presents our strategy for people matching based on color histograms and shows how tracking and recognition can be achieved. Results are presented in Section 4 and Section 5 is a conclusion.

2. People Detection

The background is estimated by computing the weighted sum between the current image in the sequence and the previous estimate of the background. The value of the weight for each pixel in the current image depends on the motion that occurred between the pixel in this image and the same pixel in the previous image. Segmentation is accomplished by applying a threshold to the absolute difference between the current image and the background estimation. Before a threshold is applied to the difference image, it is filtered using a median filter. This is done to remove any impulsive noise in the difference image. A minimum threshold is used to prevent any perturbations in the scene from causing a large number of false positive foreground pixels if the variances of the pixels become very small. Also, if there is a person passing through the scene, this person's shape as seen in the segmented image is sometimes slightly fragmented. This fragmentation may cause most person detection methods to fail. To remove as much of this fragmentation as possible a morphological closing operation is performed on this binary image. Figure 1 shows an example of silhouette obtained by following this procedure.

To detect a person, the foreground regions resulting from background subtraction are analyzed to determine if the shape of a head appears and if a body is below this detected head. To do this, local vertical peaks on the boundary of each silhouette are located using Quasi-Topological Codes. From each local peak found, the silhouette boundary is scanned in the left and right directions recording the curvature. If scanning the silhouette boundary in the left and right directions yields a convex curve (downward curve) then the corresponding local peak may represent the head of a person. 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

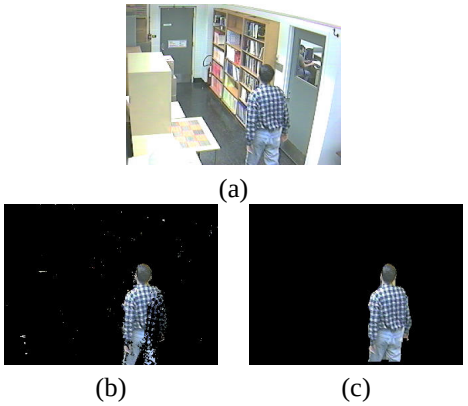


Figure 1. Example of silhouette formation. (a) A sample image. (b) The person's silhouette as initially extracted. (c) The silhouette obtained after the use of median filtering and morphological closing.

is furthest right of 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. 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.

The curvature of a silhouette boundary is determined using a square window.

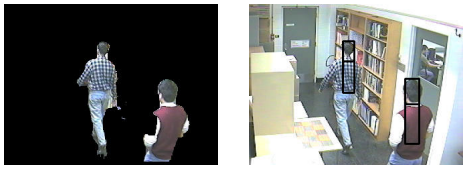
$$\left(\begin{array}{c|c} p1 & p2 \\ \hline p4 & p3 \end{array}\right)$$

$c = p1 + 2p2 + 4p3 + 8p4$

The value of c determines what the curvature of the boundary at a certain point is.

3. People Matching

In order to track and recognize people in the scene under observation, we have to be able to distinguish between different people. To distinguish between different people we decided to use clothing colour information. This will work well if the observed people wear a good variety of clothing. Also since this system monitors an area where people are passing through and therefore are seen from different



(a) (b)

Figure 2. (a) Segmented image of two people. (b) Image showing two persons and the detected regions of interest.

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 colour information is generally radially invariant. In other words, the colour of one's clothing tends to stay the same whether it is viewed from the front or from the back. Facial features are not used because the single camera, used to capture images, monitors an area that is too large to get an image of a person's face with high enough resolution to successfully perform facial recognition.

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 of this histogram 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.

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.

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.

To demonstrate how the colour histogram of a person is more similar to the colour histogram of the same person taken at a different time than colour histograms taken from different people, Figure 6 shows two images of the same person, one taken from the front and the other taken from behind, and two images, each of a different person taken from the front. Contour diagrams of the chromaticity dimensions of the colour histograms extracted as described above are shown beside their corresponding images. The two histograms taken from the same person are much more similar to each other than those taken from the other people.

To compare two histograms a measure of dissimilarity is computed using the Earth Mover's Distance, EMD, presented in [10]. 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 the 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.

If a perceptually uniform colour space is used then the EMD is a true measure of the difference between two colour histograms. Since the EMD also considers the distance between the bins in one histogram and those in the other as well as the value of each bin, it is less prone to error due to noise than methods that compare each bin in one histogram to only the corresponding bin in the other histogram.

Computation of the EMD given two 3-dimensional color histograms, H_s and H_d, proceeds as follows.

i^y_{xj}: index of the y dimension in the j'th element of S_x.

$i^{qu,v}_{xj}$: index of qu or qv in the j'th element of S_x.

c_{xj}: $(i^y_{xj}, i^{qu}_{xj}, i^{qv}_{xj})$, coordinate of the j'th element in S_x.

w_{xj}: the value of the j'th element in S_x.

$\{c_{xj}, w_{xj}\}$: the j'th element in S_x.

d^{sd}_{jk}: the distance between c_{sj} and c_{dk}.

f^{sd}_{jk}: the amount of 'earth' moved from the j'th element in S_s to the k'th element in S_d.

1. Construct the source signature, S_s, from H_s and the destination signature, S_d, from H_d, such that only bins with a value greater than a minimum value are represented in the signatures.

2. Find the value of f_{jk}^{sd}, $1 \leq j \leq n_s$ and $1 \leq k \leq n_d$ such that W is minimized.

$$W = \sum_{j=1}^{n_s} \sum_{k=1}^{n_d} d_{jk}^{sd} f_{jk}^{sd} \quad (1)$$

   (a) $f_{jk}^{sd} \geq 0$

   (b) $\sum_{k=1}^{n_d} f_{jk}^{sd} \leq w_{sj}$ and $\sum_{j=1}^{n_s} f_{jk}^{sd} \leq w_{dk}$

   (c) $\sum_{j=1}^{n_s} \sum_{k=1}^{n_d} f_{jk}^{sd} = min(\sum_j w_{sj}, \sum_k w_{dk})$

   Finding the value of f_{jk}^{sd} that minimizes W is the same as the transportation problem in linear programming and is done using the simplex method.

3. The measure of dissimilarity, D, is found.

$$D = \frac{\sum_{j=1}^{n_s} \sum_{k=1}^{n_d} d_{jk}^{sd} f_{jk}^{sd}}{\sum_{j=1}^{n_s} \sum_{k=1}^{n_d} f_{jk}^{sd}} \quad (2)$$

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 for D are considered the most likely to be the same person.

Person recognition is performed to find the correspondence between a person who has just left the room and a person who was in the room immediately before the person left the room.

3.1 Tracking People

Tracking is done by determining temporal correspondences between the people detected in the current image and those detected in previous images. The colour histogram extracted from a person detected in the current image is compared to the histogram extracted from any person detected in previous images. For each frame of the sequence, the tracking information of each person currently being tracked is updated according to the results of person detection on the current frame.

When people are detected in the current frame and at least one person is currently being tracked then the similarity between each person detected and each person currently being tracked is computed. All dissimilar detected/tracked pairs are removed from contention as possible matches. Then each person currently being tracked is updated with the information from a person detected in the current frame if the similarity measure for this pair is smaller than that of any other pair involving either of these two persons. If any person that is detected in the current frame does not get matched with any person currently being tracked then this person has probably just entered the scene and a track is initiated for this person. If a person currently being tracked has not been matched to a person detected in at least one of N consecutive frames since they were last detected then this person has probably left the scene and a decision is made on whether this person has entered or left the room based on their tracking information.

3.2 Recognizing People

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.

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.

4. Results

To measure the effectiveness of this matching technique we have captured image sequences of different people. Two sequences are captured for each person: one of the person entering the lab and one of the person leaving the lab. A certain number of persons are randomly selected and these people are assumed to be in the lab. The matching technique is then 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.

Figure 3 shows a graph of the obtained recognition rate considering different number of persons in the matching set. Figure 4 shows the values of D when each person's leaving sequence and entering sequence are compared to each other. Figure 5 shows the average values of D when each person's leaving sequence is compared to every other person's entering sequence. These two figures demonstrate how comparing a person's leaving sequence to every other person's entering sequence yields, on average, significantly higher values of D than when a person's leaving and entering sequence are compared to each other.

5. Conclusion

A method to detect, track and recognize people in indoor scenes has been presented. People are matched using

color histograms computed on a perceptually uniform color space. Our implementation of this approach was able to process approximately 2 frames per second on a Pentium II 266MHz computer. Considering the complexity of the task of human recognition and taking into account the fact that low resolution images were used, the obtained recognition rate can be considered excellent.

References

[1] C. Stauffer, W.E.L. Grimson, Adaptive Background Mixture Models for Real-time Tracking *Proceedings of CVPR*, 19990, pp. 246-252.

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

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

[4] L. M. Fuentes, S. A. Velastin, "People Tracking in Surveillance Applications". Proceedings 2nd IEEE Workshop on PETS, Kauai, Hawaii, USA, Dec. 2001.

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

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

[7] T. Gevers, A.W.M. Smeulders, "Color Constant Ratio Gradients for Image Segmentation and Similarity of Texture Objects", ISBN 0-7695-1272-0/01, IEEE, 2001.

[8] M. J. Swain, D. H. Ballard, "Color Indexing" International Journal of Computer Vision, 7:1, pp 11-32, 1991.

[9] F. Mindra, T. Moons, L. Van Gool, "Recognizing Color Patterns Irrespective of Viewpoint and Illumination" Proceedings IEEE Conf. on CVPR, vol 1, pp 368-373, June, 1999.

[10] 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 University*, 1998.

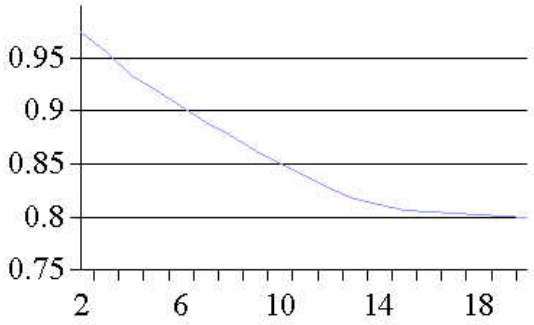


Figure 3. Graph showing recognition rate vs. number of people in the sample set.

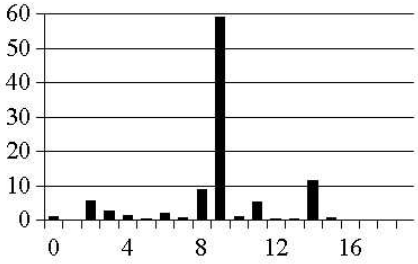


Figure 4. A graph showing the values of D when each person's leaving sequence and entering sequence are compared to each other.

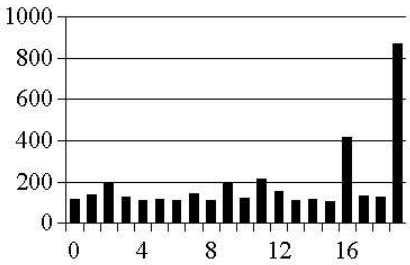


Figure 5. A graph showing the average values of D when each person's leaving sequence is compared to every other person's entering sequence.

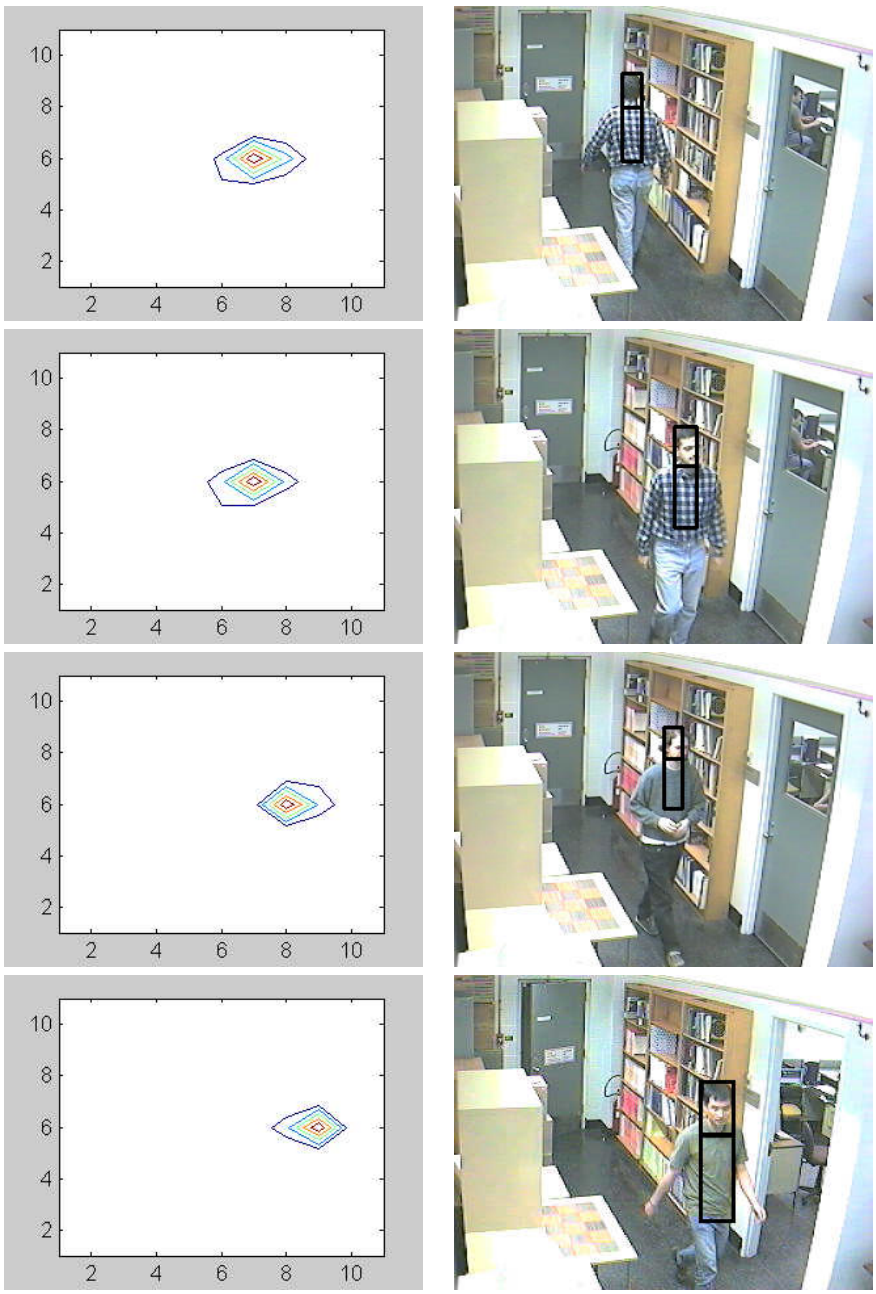


Figure 6. Images of three different people and the corresponding contour diagrams of the two chrominance dimensions.