MULTIPLE FACES TRACKING USING LOCAL STATISTICS

Sebastien Harasse, Laurent Bonnaud, Michel Desvignes

MULTIPLE FACES TRACKING USING LOCAL STATISTICS

Sebastien Harasse Laurent Bonnaud Michel Desvignes
LIS-ENSIEG
61 rue de la Houille Blanche
BP 46 38402 St. Martin d'Heres cedex France
harasse,bonnaud,desvignes@lis.inpg.fr

Abstract

Our project is to design algorithms to count people in vehicles such as buses from surveillance cameras' video streams. This article presents a method of detection and tracking of multiple faces in a video by using a model of first and second order local moments. The three essential steps of our system are the skin color modeling, the probabilistic shape model and bayesian decision, and the tracking. An iterative processus estimates the position and shape of multiple faces in images, and tracks them. Tracking updates an object history including all spatial and temporal information about this object. Location and size of these tracking object are predicted by constant speed motion analysis and learned trajectories. Results on office and buses video are promising.

1 Introduction

Estimating the number of people in a noisy environment is a central task in surveillance. A real-time count can be used to enforce the occupancy limit in a building, to manage transport traffic in real time, to actively manage city services and allocate resources for public events. Our project is to add a counting system for moving platforms such as buses, in an existing classical video recorder. Images are captured with a video camera and are analyzed to determine the number of people present. The background scene is therefore not static and vary in a large number of ways : variations in lighting levels, patterns of scene background, movements of objects that might appear or disappear in the scene. The point of view is defined by the location of the camera, in front of the people. This motivates our approach, which is to find people by finding faces.

2 Previous works

Finding people in images is a difficult task [1] due to the high variability in the appearance of people. For human detection and tracking for surveillance, various approaches have been proposed in the past years [2, 3]. We can distinguish several main classes of approaches to the problem [2].

Background subtraction [5] is often a first step to find objects of interest such as faces. Unfortunately, this approach needs a stationary background as well as interframe motion based approaches [6]. Expert systems and structural methods try to represent and recognize face by rules [7], more often describing spatial relationships between parts of a face. Finding good rules is often an extremely difficult task in an unconstrained environment.

Classical Template matching methods require the learning of several patterns of the whole face [8]. Recent works [9, 10] on template matching deal with variation in scale, pose, or shape, in the context of pedestrian detection. In their acquisition conditions, the great variability of human is dramatically reduced.

Feature based approaches extract invariant structural features from one or more images, and then classify extracted objects with statistical classifiers such as Support Vector Machines [11], neural Networks [12], probabilistic approach [13], or cascade of filters [14]. Features are designed to be invariant to some changes in illumination and pose. Several works use Harr wavalets [14], DCT [15] or local descriptors [16]. However, the largest adopted feature is skin color [2, 3]. Human skin color forms a relatively tight cluster in color spaces, even when considering darker and brighter skins. Color allows fast processing and is robust to changes in pose and illumination.

The main problem of methods base on skin color is the determination of a probability threshold to be used on each pixel for deciding which pixels correspond to the skin: this is a segmentation step which gives a binary mask for further processing. Our method solves this problem by using a Bayesian approach and reporting the decision to upper stage of processing, combining that real valued information with temporal information. The essential steps of our system are the skin color modeling, the probabilistic shape model and bayesian decision, and the tracking.

3 System overview

The system overview is illustrated by the data flow (figure 1). The detection can be summarize by steps 1,2,3 un-

less tracking is represented by step 4,5,6,.

1. step 1 is the classical bayesian detector applied to skin detection with gaussian model
2. step 2 is the computation of face probability from skin probability, face model, and prediction from previous images
3. step 3 is a decision step from face probability and a priori knowledge such as percentage of skin in face, size of face.
4. step 4 updates the history, global memory of the scene. Motion of each object is computed in this step.
5. step 5 predicts the next location of objects from motion.
6. step 6 computes and stores trajectories and also predicts face location in next image using a learning method.

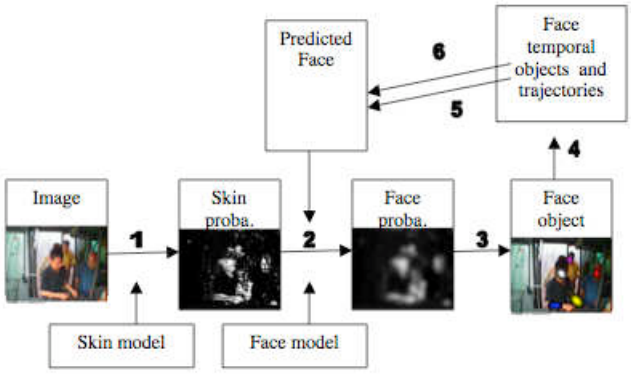

Figure 1. System overview : Data flow chart

In this system, two data levels are defined :

1. Pixel level : skin and face detection are based upon bayesian classifiers and their results are probabilities maps. Such maps are easily combined and merged to enhance or decrease probabilities values. At this level, every results will be given under this format.
2. Object level : after decisions step, objects are obtained, i.e. symbolic representation of face and human. Object tracking is a high level process on these representation and is translated in probabilities map to be merged at the pixel level during the next frame analysis.

4 Skin Model

A skin color model is needed in order to decide whether a pixel is of skin color or not. Skin chrominance is very specific, as opposed to its luminance, which has a large variability. Numerous works have been done on skin models and several color spaces have been tested [2, 3]. None of them seems to be really a better choice. In practice, intensity illumination invariant space (Normalized RGB, Lab, Yuv,..) have quite similar performances. Moreover, none of these space is color illumination invariant, which is the major variability in our application (bright sunlight, light in tunnel, sodium light in night, etc). Thus our model is defined in a chrominance color space so that skin pixels are represented in a small portion of the space. The normalized-rgb color space is used here because it is a 2D model, and because with other space color have not out preformed this one on our real data.

Since $r + g + b = 1$, only two components (r,g) are used for the model. A bidimensionnal gaussian model g_{skin} is obtained to represent skin color in the rg-space. Its parameters are learned from skin pixels from the FERET faces database.

This model is applied to an image to obtain a skin map S_I where each value is the value of our bidimensionnal gaussian model at the corresponding pixel's color. For an image I, and skin model g_{skin}, the corresponding skin map S_I is :

$$S_I(x, y) = g_{skin}(I(x, y))$$

where (x, y) is a position in the image and $I(x, y)$ is the color of I at this position, in normalized-rgb coordinates.

5 Shape Model and Bayesian Framework

5.1 Statistical Modeling

Our face detector is based on a statistical representation of the problem : a face is a skin region, parametrized by its position, shape and orientation.

Let x be a 5-dimensional random variable modeling the position and shape of a skin object, by its first and second order moments :

$$x = (\mu_x, \sigma_x)$$

with

$$\mu_x = (\mu_{x1}, \mu_{x2}), \sigma_x = \begin{bmatrix} \sigma_{x11} & \sigma_{x12} \\ \sigma_{x12} & \sigma_{x22} \end{bmatrix}$$

Our face model can be seen as an ellipse centered in μ_x with axes defined by covariance matrix σ_x. This model has been introduced in [18] for one single face tracking using color.

And let z be a random variable representing each observed image. That is to say, the realizations for z are the images where faces are to be detected. The face detection problem then involves computing the probablity density $p(x/z)$, from which we can decide where faces are likely to be in the image.

Considering a bayesian framework, the a posteriori probability density $p(x/z)$ is proportional to the product of the observation density $p(z/x)$ by the a priori density $p(x)$:

$$p(x/z) \propto p(z/x).p(x)$$

For now, x is assumed to be uniformely distributed, which means that any kind of skin colored object could be detected in the image, not restricted to faces only. $p(x)$ is constant. It is always possible to choose another distribution for x if necessary.

5.2 Observation density

The observation probability density $p(z/x)$ must now be defined. It represents the probability to observe the image z, knowing that a skin object parametrized by x is present. The number of skin objects in the image is not known, and $p(z/x)$ should allow the estimation of the number of objects and their parameters. Since random variable x is defined as the parameters of only one object, it is 5-dimensional, which is reasonable, but it doesn't allow directly the estimation of many objects. Thus $p(z/x)$ is defined so that there is a local maximum for each x corresponding to a skin object in the image.

The function chosen for the observation probability is the correlation function between the skin map S_z of image z and the bidimensional gaussian g_x parametrized by x :

$$p(z/x) \propto \int S_z(t).g_x(t)dt$$

where t is a bidimensional variable.

$p(z/x)$ has local maxima for each skin object in z, under the hypothesis that objects are well separated from each others. The proof involves approximation of S_z with a bidimensional gaussian functions mixture, and studying the behavior of $p(z/x)$ for values of x in the neighbourhood of each skin object.

Figure 2. (a) original image, (b) skin map , (c) five detected objects

5.3 Skin objects detection

From this point, there are several ways to detect the objects in our image, including the exhaustive search of local maxima in the 5-dimensional function $x \longrightarrow p(z/x)$, or sampling algorithms like Condensation [17]. We propose a method that doesn't require the calculation of $p(z/x)$ for all values.

The random variable x can be seen as two random variables μ_x and σ_x which represent the first and second order moments of the object respectively. The method proposed here estimates μ_x by using a priori information about σ_x, then estimates σ_x for each detected object, using an iterative process.

5.3.1 First order moment estimation

The detection of the first order moments μ_x of objects in the image involves an a priori estimation of σ_x. σ_m is defined as the average covariance matrix representing a face. With this assumption, the observation density becomes :

$$p(z/\mu_x, \sigma_x = \sigma_m) \propto \int S_z(t).g_{\mu_x,\sigma_m}(t)dt$$

$$p(z/\mu_x, \sigma_x = \sigma_m) \propto \int S_z(t).g_{0,\sigma_m}(t-\mu_x)dt$$

The observation density with fixed $\sigma_x = \sigma_m$ is proportianal to the 2-dimensional convolution product of S_z by a gaussian function with covariance matrix σ_m, which is an inexpensive computation. Objects' first order moments are detected by finding the local maxima of the function.

5.3.2 Iterative second order moment estimation

Suppose that an object x_0 is present in the image, with first order moment μ_{x_0}. Its second order moment σ_{x_0} must be estimated so that $p(z/x_0)$ is a local maxima.

If there is only one skin object in the image, the problem is simply resolved by computing the second order moment of the whole skin map :

$$\sigma^2_{x_0} = \int (t-\mu_{x_0})^2.S_z(t)dt$$

where t is a bidimensional variable.

Since the number of objects in the image is unknown, this method doesn't apply. Our method is to estimate σ_{x_0} by using local moments iteratively. Let W be a 2-dimensional window defined in the same space as S_z, with $\int W(t)dt = 1$. The second order local moment of S_z centered in μ_{x_0} is defined as :

$$\sigma^2_{S_z,W} = \int (t-\mu_{x_0})^2.S_z(t)W(t)dt$$

A sequence of local moments is defined as :

$$\begin{cases} \sigma_0 = 1 \\ \sigma^2_{n+1} = \sigma^2_{S_z,g(\mu_{x_0},\alpha.\sigma_n)} \end{cases} \quad (1)$$

where $g(\mu_{x_0}, \alpha.\sigma_n)$ is the bidimensional gaussian window of first and second order moments μ_{x_0} and $\alpha.\sigma_n$ respectively, with $\alpha > 2$.

Practically, the method consists in starting with a window centered in μ_{x_0} with a size smaller than the expected

object size, computing the local moments of S_z in this window, then using the result multiplied by a constant α as the next window covariance matrix. This sequence converges to the second order moment of the skin object. By using local moments, the computation of σ_{x_0} is not disturbed by the other objects in the image. The detection of multiple skin objects in the image can then be achieved. Figure 2 shows the results obtained with this method.

6 Tracking

Tracking is based upon a global structure named history, in which all information about existing and past objects and decision are stored. Tracking decision are done as soon as possible, using a prediction verification or observation scheme. When an object is present at any time, some hypothesis about its location in the next frame can be done, because motion is generally smooth. These object level hypotheses are translated at the pixel level and verified on the next frame, by comparing and merging this pixel level hypotheses with the pixel level detection (observation).

Most of the time, in our application, people pass through the left side to the right side of the camera corresponding to the previous prediction. Sometimes, people turn back and the previous prediction is no longer valid. Moreover, some trajectories are more frequent than others. A second prediction is defined to manage these situations, named learned trajectories prediction. From one location, the object location is predicted with different probabilities from all previous stored trajectories in its neighboring. Probability maps are an elegant way to manage these multiple hypothesis, which are easily combined with the motion prediction.

The observation step of skin objects tracking is tightly related to the recursive method used for the second order local moment estimation.

6.1 History

The history is a spatio temporal representation of the scene Each detected object is stored, with the following informations:

current position and motion
past positions and motions
dates of birth and death
predicted position
photometric values
shape : covariance matrix $\sum$

6.2 Prediction step

The bayesian framework is a natural way to manage several methods for a same problem, merging the results in a numeric fusion. As previously mentioned, the prediction scheme is achieved by 2 different methods, given 2 probability maps. The final prediction is a weighted sum of these 2 probabilities.

$$p_1(x) = \frac{1}{\sum_{i=1}^{2} \alpha_i} . \sum_{i=0}^{1} \alpha_i . p_i(x) \tag{2}$$

6.2.1 Motion based prediction

As motion is generally smooth, objects are predicted with a smoothed constant speed by :

$$\vec{x}_t = \vec{x}_{t-1} + \tfrac{1}{N} . \textstyle\sum_{i=1}^{N} \vec{v}_{t-i} \ \ , \ \ p_1(x) = e^{-(\vec{x}_t^T . \sigma_x . \vec{x}_t)} \tag{3}$$

If an object disappears, this prediction step is realized with the stored values unless this virtual object is predicted out of the field of view or when it has disappeared from a long time. Note that only one hypothesis is given by this method.

6.2.2 Learned trajectories prediction

As the point of view is fixed, all trajectories are learned to predict the next most probable location of an object. This is done for every position : the underlying idea is that some trajectories are more common than other and well predicted, but rare trajectories must also be predicted with lower probability. For example, in some buses, nearly all the movements are from the left side towards the right side because most of people go into the buses. But sometimes, a person goes out, and cross the field of view from the right side to the left side.

Note that for a given location, it may be possible that no objects have been detected exactly at the same place, due to real previous trajectories or due to poor segmentation accuracy. Then, the learned prediction is computed with all the previously detected neighboring trajectories. From the location x, all the movements given by the previous trajectories are reported on the current frame. Then, the probability is a mixture of gaussian, representing the object shape, weighted by decreasing distance function. All the past movements from location x are noted ($\vec{x}_k$,$\vec{v}_{x,k}$).

$$\begin{aligned} &possible\ locations\ from\ x : \vec{x}_i = \vec{x}_{t-1} + \vec{v}_{x,i} : \\ &p_2(x) = \textstyle\sum_{i=t-1}^{0} e^{-||\vec{x}_i - \vec{v}_{x,k}||^2} . e^{-\left(\frac{||\vec{x} - \vec{x}_i||}{\sigma_x}\right)^2} \end{aligned} \tag{4}$$

With this probability scheme, multiple hypotheses are managed (figure 3)until the decision step.

Figure 3. Prediction with multiple hypotheses

6.3 Observation step

The observation step chooses and corrects the predicted position and shape of the object with respect to the observed image. The same iterative algorithm than those of skin object detection is used here. The gaussian function parametrized with the most probable predicted state (position and shape) defines the window in which the first and second order local moments of the object are computed. The initial values are :

$$\begin{cases} \mu_0 = \frac{1}{card(p(x).p_{skin})} \cdot \sum_{Image} p(x).p_{skin} \\ \sigma_0 = \frac{1}{card(p(x).p_{skin})} \sum_{Image} (p(x).p_{skin} - \mu_0)^2 \end{cases} \quad (5)$$

This step is iterated by using the last computed local moments as the parameters of the gaussian window :

$$\mu_{n+1} = \mu_{S_z, g(\mu_n, \alpha.\sigma_n)} \ , \ \sigma^2_{n+1} = \sigma^2_{S_z, g(\mu_n, \alpha.\sigma_n)} \quad (6)$$

with $\mu_{S_z, g(\mu_n, \alpha.\sigma_n)}$ the first order local moment of S_z in the window $g(\mu_n, \alpha.\sigma_n)$, defined by :

$$\mu_{S_z, g(\mu_n, \alpha.\sigma_n)} = \int t.S_z(t).g(\mu_n, \alpha.\sigma_n)dt$$

In this sequence, the σ update step is the same as equation (1). This sequence converges to the first and second order moments of each object for the current image.

6.4 Occultation

One of the main feature in tracking system is the occultation problem, when people pass behind an obstacle or when people trajectories intersect. In the first case, the observation step will not find any real skin object at the predicted location (μ_0 is equal to zero in equation 5) and the object is just marked as disappeared in the history. This object is still present in the history and will be active in all step where it is necessary, such as prediction.

Figure 4. tracking example

In the second case, illustrated figure 4, the same scheme is followed : one object is marked as disappeared when the trajectories intersect, but is still present and predicted. In these cases, objects are kept in the history unless they are predicted out of the image or when they have disappeared from a long time. Moreover, the learned trajectories prediction with multiple hypotheses management, is able to track 2 persons crossing behind an obstacle.

7 Results

This tracking method has been tested under real conditions, on video streams from a transport vehicle or from an indoor office.

The office video stream has the following characteristics: 15' duration, spatial resolution 640x480 pixels, frame rate of 30 fps, and illumination conditions fairly controlled. In practice, some background objects have skin color at some moments. During this sequence, 72 persons are present in the temporal scene. Most of them (38) cross the image from left to right, 34 cross from right to left. One turns back. There are 37 intersections of trajectories. Some people have not a constant direction. In this sequence, generally more than 3 persons are visible on the same image. The bus video stream has the following characteristics: 2x3' duration, spatial resolution 702x576 pixels, frame rate of 6 fps, and illumination conditions uncontrolled. In practice, many background objects have skin color at some moments. During this sequence, 30 persons are presents in the temporal scene. 29 cross the image from left to right, 1 crosses from right to left.

Using an appropriate skin model, the detection and tracking of skin objects are efficient. figure 4 shows an example of the tracking of two faces (red and violet ellipses). One arm is also detected in the middle image (white ellipse), because it is a skin object. One part of the wall are sometimes detected as skin, when the sun lights it.

In these sequences, 90 % persons have been detected and tracked. During 2 crossings in the office video, such in figure 4, the tracking has been lost, one object been really suppressed and a new one created. The turn back is well tracked, as illustrated in figure 5.

Figure 5. Sequence with turn back tracking

Another method to validate this tracking system is the trajectories history (figure 6). Most of them are quite regular, with smooth variation. Trajectories in the bottom of the figure are arm trajectories, while faces are localized at the top of the figure. The red trajectory is a real turn back. One error can be seen with the black trajectory, in the middle of the figure : this is a part of the wall, which is skin colored. It is trapped by one person passing near this part and the detection step has included this part in the walking people face. In fact, the tracking is efficient because of good multiple prediction, if the frame rate and the time to life in the

history are suitable.

Finally, faces are counted when crossing a predefined line in the images. Counter is incremented when crossing from left to right, decremented when crossing from right to left. The accuracy of the count is 85% compared to the real manual count on the office sequence, 90 % on the transport video. Most false positives were caused by arms being counted, and non-detections face were mainly caused by bad detection of skin pixels.

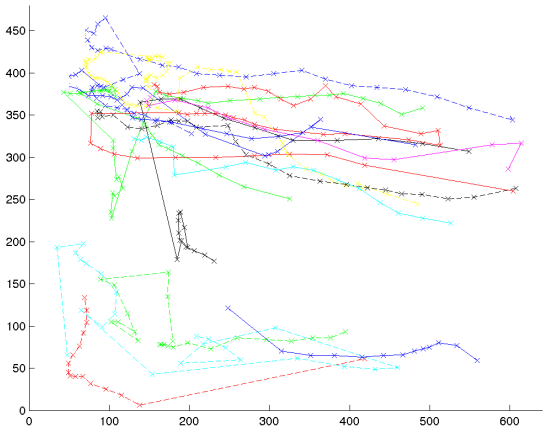

Figure 6. Detected trajectories in office video

8 Conclusion

In this paper, a multiple face tracker has been presented. Based upon the classical skin color detection, it is embedded in a bayesian framework which is an elegant way to manage several algorithms for a same task, like prediction in our tracker. The main feature is then the probability map, data structure shared by the algorithms. A spatio-temporal representation of the scene summarizes the essential information about objects. Reasoning at object level is translated at the image level, by example with the prediction probability map. Although there is no global optimization to find trajectories, multiple hypotheses are managed with an original method of prediction, based on a learning approach. Another interesting feature is the iterative local moments estimation which avoids classical thresholds, used by the skin and face detection algorithm. This approach has been applied to a counting application with promising results on out door videos.

To enhance the performance of the tracker, we are currently working on the fusion of several detection algorithm, including motion analysis and temporal difference, which can be easily integrated in our framework. Trajectories will be improved with a longer temporal and a global spatial analysis and consistency.

References

[1] S. Ioffe, D. A. Forsyth. *Probabilistic Methods for Finding People*. International Journal of Computer Vision 43(1):45-68, 2001.

[2] M.H. Yang, D. Kriegman, and N. Ahuja. *Detecting face in images: a survey*, IEEE PAMI, 24(1):34-58, 2002.

[3] Erik Hjelmas. *Face Detection: A Survey*, Computer Vision and Image Understanding, 83(3):236-274, 2001.

[4] C. Wren, A. Azarbayejani, T. Darell, A. Pentland. *Pfinder: Real-time tracking of human body*, IEEE Trans. on Pattern Analysis and Machine Intelligence, 19(7):780-785, 1997.

[5] I. Haritaoglu, D. Harwood, and L. Davis. *W4: A real-time system for detection and tracking of people and monitoring their activities*, IEEE PAMI, 22(8):809-830, 2000.

[6] Collins, Lipton, Kanade, Fujiyoshi, Duggins, Tsin, Tolliver, Enomoto, and Hasegawa. *A System for Video Surveillance and Monitoring: VSAM Final Report*, Technical report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, May, 2000.

[7] G. Yang, T.S. Huang. *Human face detection in complex background*, Pattern recognition,27(1):53, 1994.

[8] Y.H. Kwon and N. da Vitoria Lobo. *Face Detection Using Templates*,International Conference on Pattern Recognition, pp. 764-767, 1994.

[9] H. Nanda and L. Davis. *Probabilistic template based pedestrian detection in infrared videos*. IEEE Intelligent Vehicles, Versailles, France, pp 15-20, 2002,

[10] C. Stauffer and E. Grimson. *Similarity templates for detection and recognition*, Computer Vision and Pattern Recognition, pp. 221-228, Kauai, HI,. 2001.

[11] F. Xu, X. Liu, and K. Fujimura. *Pedestrian Detection and Tracking with Night Vision*, IEEE Transactions on Intelligent Transportation Systems, 5(4), 2004.

[12] H. Rowley, S. Baluja, T. Kanade. *Neural Network-Based Face Detection*, IEEE PAMI, 20(1):23-38, 1998.

[13] H. Schneiderman and T. Kanade. *Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition*, IEEE Conf. Computer Vision and Pattern Recognition, pp. 45-51, 1998.

[14] P. Viola, M. J. Jones.*Robust Real-Time Face Detection*, International Journal Computer Vision, 57(2):137-154, 2004.

[15] Z. M. Hafed, M. Levine.*Face Recognition Using the Discrete Cosine Transform*, International Journal of Computer Vision, 43 (3):167-188, 2001.

[16] V. Vogelhuber and C. Schmid. *Face Detection based on Generic Local Descriptors and Spatial Constraints*, International Conference on Pattern Recognition, Vol. 1, pp 1084-1087, 2000.

[17] M. Isard and A. Blake. *Condensation – conditional density propagation for visual tracking*, International Journal of Computer Vision 29(1):5–28, 1998.

[18] K. Schwerdt and J. L. Crowley. *Robust face tracking using color*, in Proc. of 4th International Conference on Automatic Face and Gesture Recognition, Grenoble, France, pp. 90–95, 2000.