Real-time tracking of multiple persons by Kalman filtering and face pursuit for multimedia applications

Vincent Girondel, Alice Caplier, Laurent Bonnaud

Real Time Tracking of Multiple Persons by Kalman Filtering and Face Pursuit for Multimedia Applications

Vincent Girondel, Alice Caplier, Laurent Bonnaud[1]
Laboratoire des Images et des Signaux (LIS), Institut National Polytechnique de Grenoble (INPG),
961, rue de la Houille Blanche, BP 46, 38402 Saint Martin d'Hères, France
[1]Université Pierre Mendès France (UPMF)
<Firstname.Name>@lis.inpg.fr
Tel: 33 4 76 82 62 56, Fax: 33 4 76 82 63 84

Abstract

We present an algorithm that can track multiple persons and their faces simultaneously in a video sequence, even if they are completely occluded from the camera's point of view. This algorithm is based on the detection and tracking of persons masks and their faces. Face localization uses skin detection based on color information with an adaptive thresholding. In order to handle occlusions, a Kalman filter is defined for each person that allows the prediction of the person bounding box, of the face bounding box and of its speed. In case of incomplete measurements (for instance, in case of partial occlusion), a partial Kalman filtering is done. Several results show the efficiency of this method. This algorithm allows real time processing.

1. Introduction

Human motion analysis is currently one of the most active research fields in computer vision. It deals with the detection, tracking and recognition of people, and more generally, the processing of image sequences involving humans in order to understand their behavior. For example, in smart surveillance applications, video access controls to sites, or interactive video systems (such as the European project *art.live*[1]), it is necessary to detect and track moving people passing in front of a camera in real time [5, 7]. Nevertheless, occlusion problems make the task of tracking and distinguishing people within a group much harder [3]. We propose a method using partial Kalman filtering and face pursuit to track multiple persons in real time even in the case of occlusions. In the *art.live* project, several players could play virtual games together, filmed separately by one camera each, like in Figure 1 where the person tries to catch virtual butterflies. It is now possible to make scenarii with a single camera where people can interact in the real world in addition to their interaction in the virtual world.

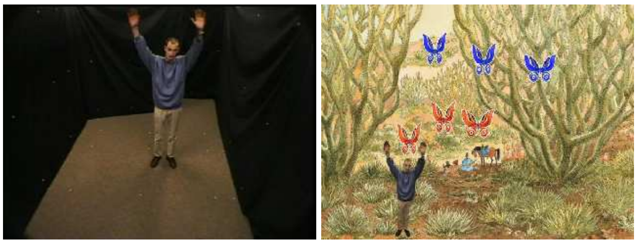

Figure 1. Example of mixed reality game. left: original image, right: mixed reality image (copyright Casterman, F. Place and project *art.live*).

Previous work

Tracking is a crucial step in human motion analysis, for it temporally links features chosen to analyze and interpret human behavior. There are numerous ways of tracking: region-based, model-based, in 2D or 3D, in single or multiple view etc. Region-based tracking, which is the paper's context, in particular, has been widely studied.

Wren et al. [7] use small blob features statistics (position and color) to track a single human in an indoor environment. The human body is, in their work, a combination of blobs representing various body parts, such as head, torso, hands and legs. The background scene and the human body are modelled with Gaussian distributions and the human body pixels are classified as belonging to particu-

[1] *art.live*: IST project 10942, ARchitecture and authoring Tools for Living Images and Video Experiments.

lar body parts using the log-likelihood measure. Tracking all the small blobs allows to track the whole body of a single human.

A study by Haritaoglu et al. [5] for a real time visual surveillance system operates on monocular greyscale or on infrared video sequences. W^4 makes no use of color cues, instead it employs a combination of shape analysis and tracking to locate people and their body parts.

McKenna et al. [6] propose an adaptive background removal algorithm that combines gradient information and color features to deal with shadows in motion segmentation. They differentiate three levels of tracking: regions, single human and human groups. They use bounding boxes for each region that can split or merge. A group is formed of several humans, each of whom is formed of several regions under geometric constraints. They manage to obtain good results of tracking multiple persons even in the case of occlusions. It would be interesting to track persons faces in addition to people.

Dockstader and Tekalp [3] present a near real time tracking method of multiple persons in a video surveillance system. The algorithm mixes coarse motion estimates, change detection information and unobservable predictions to create accurate trajectories of moving objects. This low-level features and components mixing strategy use a modified Kalman filtering mechanism. There are few system constraints but it doesn't handle complete occlusions.

Balcells Capellades et al. [1] describe a system for tracking humans and detecting human-object interactions in indoor environments. A combination of correlogram and histogram information is used to model object and human color distributions, however the particularity of human skin color is not taken into account in the color model. The models are built on the fly and used to track people on a frame by frame basis. The system is able to detect when people merge into groups and segment them during occlusion.

Our approach

The aim of the work presented here is a way to solve the difficult problem of tracking multiple persons in real time and handling partial or even complete occlusions between them. In this paper, the assumptions are the following:

1. Indoor environment filmed by **one static camera**,

2. Each person enters the scene **alone** with the **face at least partially visible**.

There are three contributions in this paper. First, a very fast and simple tracking method based on the computation of the overlap of bounding boxes is described in section 2.2 [2, 6, 7]. Then, a robust and adaptive skin detection method in the YC_bC_r color space based on thresholding in the C_bC_r plane is presented in section 2.3 [1, 4, 5]. In section 3, we propose an overall tracking method which uses the combination of partial Kalman filtering and face pursuit to track in real time people, even in the case of partial or complete occlusion problems [3].

2. Preprocessing steps

2.1. Person segmentation

The first step of this work is the segmentation of each person. It is performed by a **background removal algorithm** which computes the difference between the current image and a background reference image updated over time [2]. A Markovian relaxation is done in order to obtain a regularized segmentation mask for each person. This segmentation is followed by morphological operations, connex components labeling and yields to the computation of video objects parameters (rectangular bounding box, gravity center, surface etc.). From now on, the term box is equivalent to rectangular bounding box.

2.2. Temporal tracking

The temporal tracking between detected boxes on two consecutive images results from the combination of a forward tracking phase and a backward tracking phase. For the forward tracking phase, we look for the successor(s) of each person detected at time $t-1$ by computing the overlap surface between its box and all the boxes detected at time t. In case of multiple successors, they are sorted by decreasing overlap surface (the most probable successor is supposed to be the one with the greatest overlap surface). For the backward tracking phase, the procedure is similar: we look for the predecessor(s) of each person detected at time t. Considering a person P detected at time t: if P's most probable predecessor has P as most probable successor, a temporal link is established between both boxes. If not, we look in the sorted lists of predecessors and successors until a correspondence is found, which is always possible if P's box has at least one predecessor. If this is not the case, P is a new person.

This basic tracking is **very fast to perform** and allows:

1. Segmentation problems **correction**: If one object of human size is split in several little parts, we can merge them back by comparing the surfaces of these several successors boxes to the human size and correct the segmentation. If a group of people is split into several persons, nothing is done because each of them is of human size in surface.

2. Occlusion problems **detection**: If several little parts (smaller in size than a human one) merge into a bigger object, nothing is done. If several persons merge into a group, this group will have several predecessors and a size twice or at least greater than a human one so we may have occlusion problems.

2.3. Face localization

Face localization is carried out in two steps: skin detection, and face and hands identification.

Skin detection is performed on all the pixels of segmentation masks detected as persons. Only the pixels inside persons boxes are processed and not the background pixels. Skin pixels are detected in the YC_bC_r color space by thresholding on the C_b and C_r components with the initial following thresholds, learned from a sample of skin images database with our camera: $C_b \in [90; 130]$, $C_r \in [130; 180]$. Indeed it gives a detection quality similar to that of other color spaces, with a lower computational cost [4]. In order to make the skin pixels detection more robust with respect to each detected person, to the acquisition system and to illumination conditions, thresholds are adapted over time but restricted inside the initial square. Initial detection intervals are gradually translated and reduced (one color unit per frame) towards the C_b and C_r means of detected skin pixels (15x15 color units minimum size for the C_bC_r detection box in the C_bC_r plane). On Figure 2 are drawn the skin database pixels in the C_bC_r plane, the big black square corresponds to the initial thresholds for skin detection and the other little squares adapted thresholds for different persons.

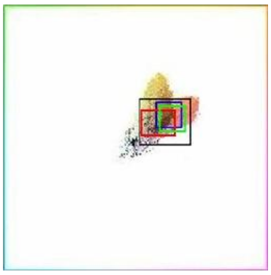

Figure 2. Skin database pixels in C_bC_r plane.
big black square: initial thresholds,
little colored squares: adapted thresholds

Then, criteria related to temporal tracking and to human morphology (size, position in the bounding box etc.) are used in order to identify the face and hands among all the connex components labelled as video skin patches [4]. Faces and hands bounding boxes are also computed. Although the algorithm can detect and track the face and hands efficiently for a single person, in the case of multiple persons with occlusion problems, hands detection and tracking are not as reliable as the face detection and tracking. Hands are smaller, move faster and have a less steady movement than the face. Therefore we only consider the person and face boxes for multiple persons tracking.

3. Kalman filtering

The Kalman filter is a well-known optimal and recursive signal processing algorithm for parameters estimation. With respect to a given model of parameters evolution, it computes the predictions and adds the information coming from the measurements in an optimal way to produce *a posteriori* estimations of the parameters. Since there is no command vector, here are the state vector prediction and filtering equations:

$$\begin{aligned} \underline{\hat{x}}_{\,t/t-1} &= A_{t-1}\underline{\tilde{x}}_{\,t-1/t-1} \\ \underline{\tilde{x}}_{\,t/t} &= \underline{\hat{x}}_{\,t/t-1} + G_t(\underline{m}_{\,t} - C_t\underline{\hat{x}}_{\,t/t-1}) \end{aligned}$$

At time $t-1$: $\underline{\hat{x}}_{\,t/t-1}$ is the predicted state vector, A_{t-1} the model's evolution matrix, $\underline{\tilde{x}}_{\,t/t}$ the *a posteriori* estimated state vector, G_t the Kalman gain matrix, $\underline{m}_{\,t}$ the measurements vector, C_t the measurements matrix.

We use a Kalman filter for each new detected person. With regard to the real time constraint, simple choices have to be made: the global movement of a person is supposed to be the same as the face movement. Associated with a constant speed evolution model, this leads to a state vector of ten components for each Kalman filter: the rectangular bounding boxes of the person and face (four coordinates each) plus two components for face speed:
$\underline{x}^T = (x_{pl}, x_{pr}, y_{pt}, y_{pb}, x_{fl}, x_{fr}, y_{ft}, y_{fb}, v_x, v_y)$. The indexes p and f respectively stand for the person and face box, l, r, t and b respectively stand for left, right, top and bottom coordinate of a box. v_x and v_y are the two components for the 2D apparent face speed. We have therefore the following Kalman filter's evolution matrix:

$$A_t = A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

In our case, $C_t = I_d$. All these choices yield to the following equations presented in computation order where $P_{t+1/t}$ is the predicted covariance matrix, Q the model's noise matrix, R the measurements noise matrix and $P_{t/t}$ the *a posteriori* estimated covariance matrix. We have also the following definitions: $R = Diag(\sigma_i)$ and $Q = Diag(\sigma_i^{'})$.

$$\begin{aligned} G_t &= P_{t/t-1}(R + P_{t/t-1})^{-1} && (1) \\ P_{t/t} &= (I_d - G_t)P_{t/t-1} && (2) \\ \underline{\tilde{x}}_{\,t/t} &= (I_d - G_t)\underline{\hat{x}}_{\,t/t-1} + G_t\underline{m}_{\,t} && (3) \\ \underline{\hat{x}}_{\,t+1/t} &= A\underline{\tilde{x}}_{\,t/t} && (4) \\ P_{t+1/t} &= AP_{t/t}A^T + Q && (5) \end{aligned}$$

Initial conditions:

$$\begin{aligned} \underline{\hat{x}}_{\,0/-1} &= \underline{x}_{\,0} \\ P_{0/-1} &= P_0 = \lambda I_d \end{aligned}$$

3.1. Motion estimation

For each face detected at time $t-1$ by the skin detection method, we compute a speed estimation by classical methods such as block-matching or histogram matching. Kalman filter's evolution matrix, A, expresses the fact that the same speed is used for the evolution of person and face boxes. Measurements come from boxes provided by the segmentation step, face boxes obtained by the skin detection step and face speed estimations. Segmentation provides boxes that can contain several persons whereas the Kalman state vector (and therefore the Kalman person box) is defined for a single person. Three different person bounding boxes are associated to each person: the box given by the segmentation step that may contain other persons in the case of a merge, the Kalman predicted person box and the Kalman *a posteriori* estimated person box.

3.2. Individual tracking mode

Individual tracking mode is selected when all person box measurements are available *i.e.* there are neither occlusion nor merge problems. The segmentation box contains a single person. More generally, in individual tracking mode:

1. The person's face is detected at time t (all face box measurements are available) and

2. The person's face has been detected at time $t-1$ (face speed estimation measurements are available).

If **1.** or **2.** measurements are not available (face localization step has failed respectively at time t or $t-1$), we are in a particular case of partial Kalman filtering. Unavailable measurements are replaced by the Kalman predicted values. The Kalman filtering is then done for all state vector components, including those that have their measurements replaced by Kalman predicted values. Equations (1) to (5) are used.

3.3. Partial Kalman filtering and predictive modes

Partial Kalman filtering mode is selected when:

1. Some person box measurements are not available *i.e.* there are occlusion and/or merge problems and

2. The Kalman *a posteriori* estimated person box contains a unique face.

Available person box measurements can serve for different persons boxes. For example, on image 203 on Figure 3, if two persons have just merged (hands touching), we have only four measurements available (instead of eight) that can be used as estimations for the persons boxes. Each Kalman *a posteriori* person box contains a unique face. The left, top and bottom side measurements of the segmentation box containing both persons will serve as measurements for the Kalman predicted person box on the left side. The right and bottom side measurements will serve as measurements for the Kalman predicted person box on the right side.

The attribution of available measurements is decided considering the different boxes centers and sides coordinates. The comparison is performed between the segmentation boxes and the Kalman predicted person boxes. For the person box, we generally have two or three available measurements (up and/or down side(s) and one side measurements).

For the face box or the face speed, we have either all measurements available or none. If we have none, Kalman predicted values replace the measurements. Equations (1) to (5) are used as long as Kalman *a posteriori* estimated person box contains a unique face. If there is more than one face overlapped by the Kalman *a posteriori* estimated person box, the Kalman filter works in predictive mode since the face localization step could provide inaccurate positions.

Kalman predictive mode is selected when:

1. Some person box measurements are not available *i.e.* there are occlusion and/or merge problems and

2. The Kalman *a posteriori* estimated person box contains more than one face.

No measurements are taken into account. All the state vector components are predicted according to the last face speed estimation. Only equations (4) and (5) are used. The Kalman filter works in predictive mode until a unique face is again detected in the Kalman *a posteriori* estimated person box.

4. Results

Thanks to the adaptation process, skin detection is robust enough to provide good results even on backgrounds whose color is like skin. We manage to detect and track faces and hands on a yellow/brown background. Indeed, detection intervals form a compact square in the C_bC_r plane.

Figure 3 illustrates successful tracking results on a video sequence with an easy background in which a person is completely occluded. Segmented and tracked persons are visible on the original images of the sequences. Segmentation boxes are drawn in white lines, Kalman *a posteriori* estimated boxes (person and face) in color dashed lines. In the first sequence, images 200, 228 and 231 show an individual tracking with all measurements available for the Kalman filter before the merge (image 203) and after the split (image 228). Images 212, 213 and 219 illustrate the Kalman filter in a predictive mode when one face is occluded. Images 203 and 221 (just before the split) illustrate the partial Kalman filtering.

The second sequence, on Figure 4, also shows some successful results although the more difficult and general background induces uncorrected segmentation problems.

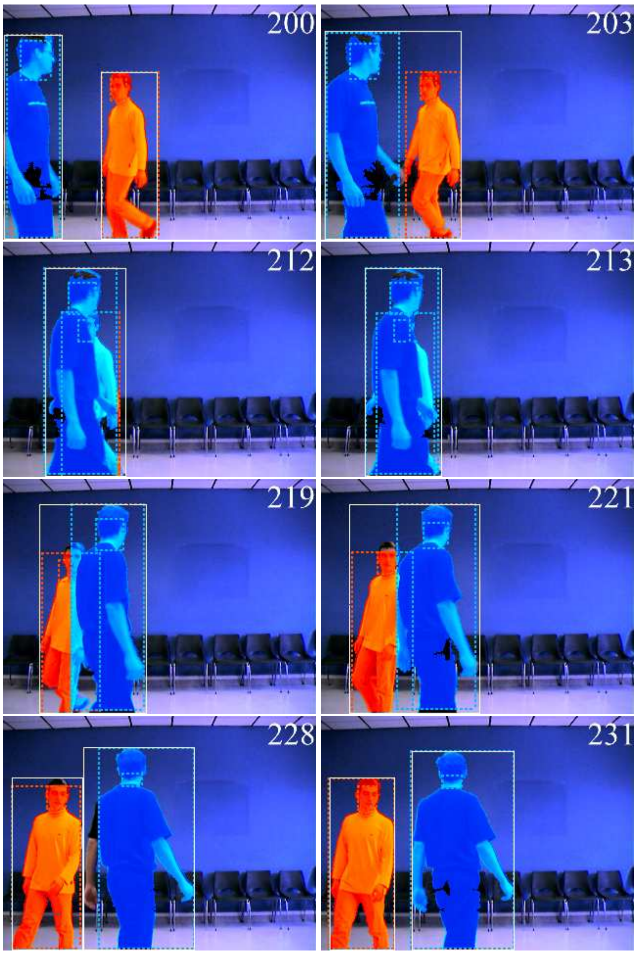

Figure 3. Results of a test sequence (easy background).

The whole sequences are available in *MPEG* at the URL: `www.lis.inpg.fr/pages_perso/bonnaud/SSIAI/`.

Computing time

Video sequences were acquired in the YC_bC_r $4:2:0$ format at 30 fps and in 640×480 resolution. The Figure 3 and Figure 4 results were obtained at a frame rate of approximatively 10 fps on a low-end PC running at 1.8 GHz. Considering the fact that the C++ code is not optimized and that classical resolutions of sequences are 320×240 or even 160×120, real time processing could be easily achieved.

5. Conclusion and perspectives

We have presented in this paper a fast and efficient algorithm based on the combination of partial Kalman filtering and face pursuit in order to track multiple persons even under occlusions. This method can be used for indoor video sequences. Further work will address color modeling of a person's body parts and eventually behavior recognition.

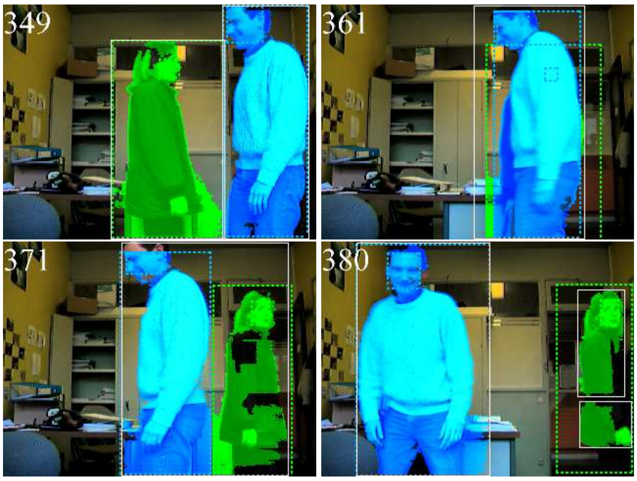

Figure 4. Results of a test sequence (difficult background).

References

[1] M. B. Capellades, D. Doermann, D. DeMenthon, and R. Chellappa. An appearance based approach for human and object tracking. In *IEEE International Conference on Image Processing*, September 2003.

[2] A. Caplier, L. Bonnaud, and J.-M. Chassery. Robust fast extraction of video objects combining frame differences and adaptative reference image. In *IEEE International Conference on Image Processing*, September 2001.

[3] S. L. Dockstader and A. M. Tekalp. On the tracking of articulated and occluded video object motion. *Real Time Imaging*, 7(5):415–432, October 2001.

[4] V. Girondel, A. Caplier, and L. Bonnaud. Hands detection and tracking for interactive multimedia applications. In *International Conference on Computer Vision and Graphics*, pages 282–287, September 2002.

[5] I. Haritaoglu, D. Harwood, and L. Davis. Who, when, where, what: A real time system for detecting and tracking people. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 222–227, April 1998.

[6] S. J. McKenna, S. Jabri, Z. Duricand, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision Image Understanding*, 80:42–56, 2000.

[7] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-time tracking of the human body. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 19(7), pages 780–785, July 1997.