Real-time face detection and tracking for vending machine data management

Benoit MARTIN[1], Julien MAROT[2], Salah BOURENNANE[2]

[1]IntuiSense Technologies,
30 avenue du Château de Jouques
13420 Gemenos, France

[2]CNRS, Aix Marseille Université, Centrale Marseille, Inst. Fresnel,
52 av. Normanide 13397 Marseille, France
benoit.martin@intui-sense.com, julien.marot@fresnel.fr, salah.bourennane@fresnel.fr

Résumé – Habituellement, la détection et la reconnaissance de visage sont réalisées dans un environnement idéal. De plus, on recommande souvent une caméra 3D ou haute résolution, aux dépens du temps de calcul. Nous proposons un algorithme temps réel qui inclut la détection de visage, le suivi, et l'identification. Ceci pour une application de mesure d'audience au niveau de distributeurs de boissons automatiques. Les conditions de travail sont difficiles et incontrôlées, et les contraintes de temps réel et de coût sont fortes. Nous atteignons dans ce contexte un taux de bonne détection de 80%.

Abstract – Usually, face detection and face recognition studies are performed in an ideal environment. Moreover, to do face recognition, it is highly recommended to use a high quality 3D camera, at the expense of elevated cost and computational load. We propose an algorithm including face detection, tracking, and identification, to managing video data coming from a vending machine. The acquisition conditions are uncontrollable, and the real-time and cost constraints are harsh. In this context, we reach a good recognition rate of 80%.

1 Introduction

In this paper, we aim at performing data management from videos acquired by a low-cost camera included in a vending machine. Our final goal is two-fold : firstly to determine the number of persons which were present in the field of view of the camera during a certain time-lapse ; secondly to select with the highest probability and following a visual intuition, the current user of the machine.

Face detection and tracking has been a long standing problem in computer vision.

This paper is related to prior work in both face detection and object tracking fields :

In [1], a review of face detection methods distinguishes two wide categories : deformable parts-model (DPM), and rigid templates. DPM exhibit an elevated computational load [1], which is not suitable for the goals of this work. Rigid templates are for instance based on Haar-like features [2], or local binary patterns [3].

To reach both high detection rates and low false alarm rates, the the Viola-Jones detector combines multiple classifiers, using either a parallel architecture [4] or a cascade architecture [5]. In these papers, three main ideas make this face detector run in real-time : firstly, a new image representation called the "Integral image" is introduced, which allows the features used by a detector to be computed very quickly. Secondly, a learning algorithm, based on AdaBoost, selects a small number of critical visual features. Thirdly, a cascade structure permits to reject the majority of candidate sub-windows in the image in early stages of the detector.

In [6] multi-object tracking is performed. The authors propose a criterion denoted 'multi-object tracking accuracy' (MOTA) to study the performances of their algorithm. However, they restrict their study to pedestrians.

We notice that, in the previously cited references, face detection or characterization is performed only in rather good, or even optimal conditions. However, in the real-world context considered in this paper, people should be detected and tracked from one frame to another even when they rotate slightly their face from left to right or top to bottom, or if the illumination condition evolve during their movement.

The main contribution of this paper is as follows :

We propose image processing algorithms to manage video data coming from a vending machine. A low-cost camera is placed at the top of the machine, and associated with a PC with limited resources. The novel-most aspects of this work are the following : our algorithms work in real-time to determine the current number of persons in the field of view of the camera and the person with the highest probability of being the current user of the vending machine. We propose an evaluation of the algorithm on a reference video database [7], which, to the best of our knowledge, exhibits such difficult conditions that no tra-

cking but only face recognition (see for instance [8]) algorithm has been performed on it.
The paper outline is the following :
In section 2, we present the hardware, the operating system and the software we are working on. In section 3 we detail the proposed software for people detection and tracking, and for the identification of the current user of the Vending machine. Section 4 presents visual and numerical results of people detection and user identification.

2 Hardware and software materials

In this section, we describe the technical context, *i.e.* the aimed hardware and the software tools.

2.1 Hardware

The software is meant to be integrated on a Vending Machine equivalent to those which can be found on highway's stations. This machine is equipped with a little 2D camera which has a vertical *Field Of View (FOV)* of 90° and is linked to the machine's computer via an USB connection. This camera is located in the middle above the machine's screen with an angle which is wide enough to permit it to see people with a height comprised between 1m55 and 2m. The video output of the camera has a resolution of 1240×720 pixels.
The machine's computer is running with a Windows 7 Embedded Operating System and equipped with a Celeron processor @2GHz and a RAM of 4GB.

2.2 Operating system and software

The software are developed using the C++ language and the free computer vision library OpenCV in its 3.0 version. To achieve the lowest possible computing time and for cross-platform purpose, the Boost library is also used in its 1.59 version.
On the one hand, we developp and test our software on a personnal computer running a Windows 8.1 Operating System and equipped with a Intel Core i5-4460 CPU @3.20GHz and a RAM of 8GB. Afterwards, the software is run on the aimed computer described previously in subsection 2.1.

3 Proposed algorithm

Figure 1 presents the overall structure of our algorithm, which improves the detection/tracking rate while requiring a low computational load and a reduced memory space. The algorithm is separated in three main parts, *i.e.* Face Detection, Face Tracking and Data Management.
Before processing the frame t, the software should be aware of the position of the detections found on the previous frame $t-1$ and each of these detections has been given a grid of keypoints.

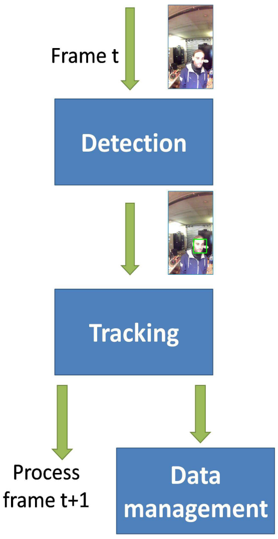

FIGURE 1 – Algorithm's Workflow

3.1 Face detection

In a simplified but also general manner, the visual aspect of a face is as follows : the eye region is darker than the upper-cheeks, the nose bridge region is brighter than the eyes, and the forehead is brighter than the hair. Hence, they appear as a combination of black and white rectangles. These rectangles are mathematically represented by the Haar features, used in the the Viola-Jones detector. The detector used has been trained using 8,000 training samples, of which 5,000 are face images and 3,000 non-face images [9]. This detector is composed of 25 classifiers. The Viola-Jones method complexity is approximately $O(n)$ with n being the number of pixels in the studied image[10]. Therefore, to lower the computational cost of the algorithm, we only apply the Viola-Jones detector on the whole picture once every 5 frames. On the 4 frames between two complete detection, we only apply the Viola-Jones detector on *Regions Of Interest (ROI)* chosen accordingly with the previous detections' location.

3.2 Tracking

The tracking part is meant to link the detections found in the current frame t to the detections found in the previous frame $t-1$. Indeed, a required piece of information is the presence time of each person during some time-lapse. Therefore, the software must be able to detect a person appearance and disappearance in the field of view. Hence the necessity of an multi-object tracking algorithm.
The used method is similar to the KLT algorithm [11]. Take the current detection i. Its position is compared to the previous detection' to compute the overlapping area. If the overlapping area is considered big enough with the previous detection j, then they are consider linked on a spatio-temporal point of view. therefore, the algorithm assimilates the current detection i as the new position of the previous detection j. Then, for each pre-

vious detection which has not been found in the current frame t with this overlapping area criteria, the detection's keypoints and the Lucas-Kanade Optical Flow [12] are used to find the detection back in the current frame t. If the detection is still not found, then the detected person is considered as disappeared from the field of view.

3.3 Data management : User identification

The data management part has two meanings. The first one is to update and save the detections' information, *i.e.* appearance time and disappearance time, in an output file. This output file could then be used to study the average presence time of the detected people or even to search the time of the day with the biggest, or the smallest, number of people present in the field of view. The second meaning, and the most important one, is to determine which person in the video is the current user of the vending machine. To do this, we use Eq. (1).

$$P(D = user) = A \times B \times C \tag{1}$$

with

$$A = \left(1 - \frac{D_x - \frac{w_{tot}}{2}}{\frac{w_{tot}}{2}}\right) \tag{2}$$

and

$$B = \left(1 - \frac{D_y - \left(h_{tot} - \frac{\bar{D_h}}{2}\right)}{h_{tot} - \frac{\bar{D_h}}{2}}\right) \tag{3}$$

and

$$C = \left(\frac{D_{area}}{Area_{max}}\right) \tag{4}$$

Where D is the studied detection, (D_x, D_y) are the detection's center coordinates, D_{area} is the detection's area, h_{tot} and w_{tot} are respectively the height and the width of the total image, $\bar{D_h}$ is the height of an average detection and $Area_{max}$ is the maximum area a detection can have. The probability P is the most elevated when the detected face is the biggest, and at the bottom-center of the FOV. Thanks to this method, the software is able to determine which person in the video is the current user with as much precision as the human eye.

4 Results

Our solution has been tested on both the ChokePoint Dataset[7] and a homemade video saved directly from the Vending machine's camera. The ChokePoint dataset is very appropriate to test the abilities of our proposed methods in a real-world environment, with a large number of persons in the scene, and varying illumination conditions. It has also been used in [8], to analyze the performances of face recognition methods. To the best of our knowledge, it has not been used to test people detection and tracking algorithms yet. Some qualitative results are available on Figure 2, with (a), (b) and (c) being the output of our solution on our homemade video and (d), (e) and (f) the output on a video from the ChokePoint Dataset.

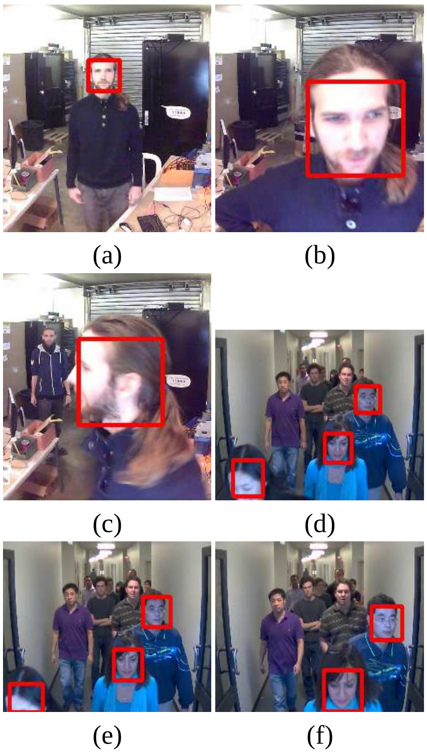

FIGURE 2 – Qualitative results : detection and tracking

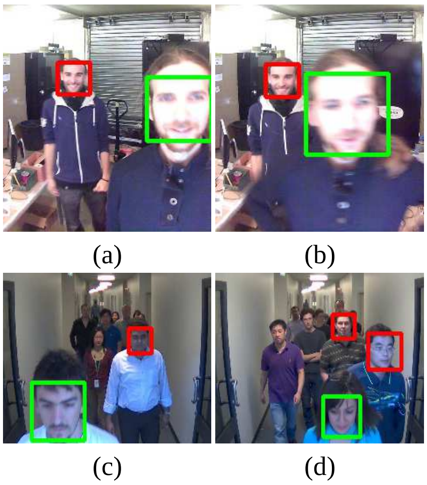

FIGURE 3 – Qualitative results : user selection

The Multi-object tracking accuracy (MOTA, see Eq. (5)) [13] is used to evaluate the performances of the proposed method.

$$MOTA = 1 - \left(\frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}\right) \tag{5}$$

where m_t is the number of misses for frame t, fp_t is the number of false positive for frame t, mme_t is the number of mismatch errors for frame t, and g_t is the number of objects present at frame t.

The quantitative results of the proposed method on the ChokePoint dataset are shown on Table 1. A particular interest should be shown regarding the sequences P2E_S5_C21 and P2L_S5_C21

Sequence	Output	Ground truth	MOTA
P1E_S1_C1	24	26	92.31%
P1E_S2_C1	21	25	84.00%
P1L_S2_C1	17	22	77.27%
P1L_S3_C3	21	24	75.00%
P2E_S3_C11	21	24	87.50%
P2E_S5_C21	19	24	75.00%
P2L_S2_C21	25	25	92.00%
P2L_S4_C21	22	25	88.00%
P2L_S5_C21	24	25	96.00%

TABLE 1 – Quantitative results on the ChokePoint dataset

as they are the most challenging videos regarding multi-object tracking. Indeed, in these two videos, all the people on the video move towards the camera together.

5 Conclusion

This work is performed in the context of intelligent industrial vision systems. We use Viola-Jones face detector, which interprets the visual aspect of faces as Haar features ; we associate the detected faces from one frame to the following through an overlapping surface criterion, and taking advantage of the optical flow computed on keypoints. Finally, we identify the user of the machine, through some assumptions on the visual aspect and the position of his face in the field of view. We exemplify the proposed method on homemade videos and on a reference database : experiments have shown that the proposed method yields elevated tracking accuracy values, an average MOTA of 85%, on videos with around 20 people to be detected. Moreover, the proposed solution also worked on a live stream with a frame rate of 20 fps and a CPU load of 40% on the Celeron CPU used on the vending machine. Therefore, the method can work in a real-time disposition.

Références

[1] Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang, "A survey on face detection in the wild : Past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1–24, Sept. 2015.

[2] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, 2001, vol. 1, pp. I–511–I–518 vol.1.

[3] Yan Zhang and Caijian Hua, "Driver fatigue recognition based on facial expression analysis using local binary patterns," *Optik - International Journal for Light and Electron Optics*, vol. 126, no. 23, pp. 4501–4505, Dec. 2015.

[4] Paul Viola and Michael J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137 – 154, 2004.

[5] Paul Viola, Michael J Jones, and Daniel Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[6] Shu Tian, Fei Yuan, and Gui-Song Xia, "Multi-object tracking with inter-feedback between detection and tracking," *Neurocomputing*, vol. 171, pp. 768 – 780, 2016.

[7] Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, and Brian C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2011, pp. 81–88, IEEE.

[8] L. An, B. Bhanu, and S. Yang, "Boosting face recognition in real-world surveillance videos," in *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, Sept 2012, pp. 270–275.

[9] Rainer Lienhart, Alexander Kuranov, and Vadim Pisarevsky, *Pattern Recognition : 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings*, chapter Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection, pp. 297–304, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[10] Zsolt T. Kardkovacs, Zsombor Paroczi, Endre Varga, Adam Siegler, and Peter Lucz, "Real-time traffic sign recognition system," in *Cognitive Infocommunications (CogInfoCom), 2011 2nd International Conference on*. pp. 1–5, IEEE.

[11] Jianbo Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, Jun 1994, pp. 593–600.

[12] Bruce D. Lucas, Takeo Kanade, and others, "An iterative image registration technique with an application to stereo vision.," in *IJCAI*, vol. 81, pp. 674–679.

[13] Keni Bernardin, Alexander Elbs, and Rainer Stiefelhagen, "Multiple object tracking performance metrics and evaluation in a smart room environment," in *Sixth IEEE International Workshop on Visual Surveillance, in conjunction with ECCV*. vol. 90, p. 91, Citeseer.