Human Automatic Tracking in Outdoor Scenes

Patrick Marques Ciarelli, Evandro O. T. Salles and Elias Oliveira
Department of Electrical Engineering
Universidade Federal do Espírito Santo
Vitória, ES, Brazil
pciarelli@lcad.inf.ufes.br, evandro@ele.ufes.br, elias@lcad.inf.ufes.br

Abstract— The use of cameras for video-based security systems is becoming more and more common. The identification of non-conventional events through cameras, such as crimes and accidents, is a powerful application of image processing in real life. In this paper, we propose an automatic approach for human tracking in outdoor environment using Kalman filter and Camshift. Furthermore, we propose a method to find the face of the person in video sequence. We have carried out a series of experiments on some video sequences and the results achieved are encouraging.

Keywords-Kalman filter;Camshift;optical flow; tracking.

I. INTRODUCTION

With the technological advancement, the cameras are becoming more and more robust, cheap, with small size and weight and presenting better image quality. These characteristics motivate their use in different applications, where they are employed to record several videos. The task to identify events in video scenes is relatively new and it has been growing the interest to employ this approach in real life. Nowadays, one of the greatest applications of cameras and identification of events is in video-based security systems. These systems are applied to detection of crime scenes, accidents, unlawful acts, non-conventional events and identify people, such as suspects and missing people [1].

However, it is important to run this activity automatically, without human interference or as minimal as possible, in order to avoid human fail situations. Another problem is the presence of partial or total occlusion of the target, which is still a challenge to be solved. The tracking task becomes more difficult in outdoor environments, where illumination conditions and the background of the scene may vary continuously.

In order to face these problems, we present in this paper an approach based on Camshift and Kalman filter to track people in outdoor environment in the presence of occlusion. However, we employ a Camshift version proposed in [2] because it is more robust in outdoor environments. To make the task of automatic tracking, we have tested three approaches to capture the initial information about the target: two based on optical flow and another one based on frame subtraction. Furthermore, we propose a method, based on our approach, to find the face of the person in video sequence. We carried out a series of experiments with our proposed technique and the results obtained are encouraging.

This work is organized in the following structure. We describe the version of Camshift used in our approach in Section II. In Section III, we discuss the integration between Camshift and Kalman filter. The procedures to segment and find the human face are mentioned in Section VI. Section V defines how our experiments were performed and the achieved results. Finally, we present our conclusions and we indicate some future directions for this research in Section VI.

II. MODIFIED CAMSHIFT

Camshift presented in [2] has the following steps:

1. Choose the initial size and localization of the search window on the target;
2. Based on the target color, compute the color probability distribution of the region, centered at the search window with an area slightly larger than the window;
3. Compute the moment of order zero and the centroid location of the window to convergence or for a number of iterations;
4. Update the target color and compute the percentage of relevant pixels in the search window;
5. In the next frame, the search window is centered in the location found in the step 3, the window size is obtained from zero order moment and the shape of the window is obtained from the relevant pixels in the window. Go to step 2.

Two approaches are common for finding a target in a scene: look for it in entire image or into a small window (search window). Camshift uses the second one [7].

The probability distribution image may be obtained using any method that associates for each pixel value a probability value. A common method that approximates the probability is the Histogram Back-Projection, introduced by Swain and Ballard in [3]. In this method, initially, it is computed the histogram of the region of interest. Then, the histogram bin values are rescaled from [0, $max(bin)$] to [0, 255], where $max(bin)$ is the largest bin value in the histogram. Finally, the histogram bin values are associated with the correspondent pixel values.

However, in this proposed version of Camshift, the histogram is obtained from the search window using a pyramidal weight scheme, as it is shown in Fig. 1. Pixels in the region P have weight p, whereas pixels in the region Q have weight q, where $p > q$. Pixels outside of the region Q have weight equal to zero.

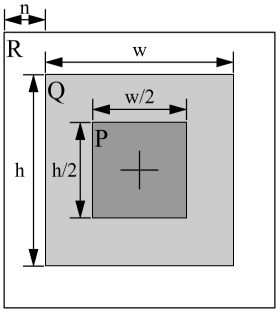

Figure 1. Pyramidal weight scheme.

The dimensions of each region are also shown in Fig. 1, where *w* and *h* are the window dimensions, and *n* is a constant number of pixels added to window size.

Since a search window is given and a color probability distribution is calculated, the next step is to compute the moments of order zero and one within the search window. This process continues until the convergence of the window position or for a number of iterations. Let $I(x, y)$ be the intensity of the discrete probability image at (x, y), the moments can be calculated by (1), (2) and (3) as follows [7]

$$M_{00} = \sum_x \sum_y I(x, y), \tag{1}$$

$$M_{10} = \sum_x \sum_y xI(x, y), \tag{2}$$

$$M_{01} = \sum_x \sum_y yI(x, y). \tag{3}$$

The new location of the centroid of the search window (x_c, y_c) is computed by

$$x_c = M_{00}/M_{10}\,; y_c = M_{00}/M_{01}. \tag{4}$$

Although in classical Camshift has been used only one channel of color space to track the target, this approach may be disastrous to track objects in uncontrolled environments. The reason is that the use of a single channel may bring difficulties for object tracking where the assumption of single channel cannot be made, such as multiple objects in a scene and places where the used channel cannot distinguish the objects from the background [2]. These problems aggravate in outdoor environments, where there are modifications on the lighting conditions and large variation on the color channel of the scene. Therefore, we propose a version of Camshift that uses more than one channel.

Let I_1 and I_2 be the probabilities distribution image for two different channels, they can be combined by

$$I(x, y) = (I_1(x, y) I_2(x, y))/255. \tag{5}$$

We have chosen to work just with two channels: hue and saturation of the HSV color space. We avoid using the value channel because it is very sensitive to illumination conditions. For a range of values of illumination, the hue does not change with illumination, and the variation in the saturation is smaller than in the value channel [4].

Then, whenever the window location converges, it is updated the histogram. The updating of histogram is used to reinforce the histogram of the target color, and thus to fit better to the target. Let $hist_t$ and $hist_{new}$ be the current histogram and the histogram computed in the current frame, and α a constant, the updated histogram $hist_{t+1}$ is reached by

$$hist_{t+1} = (1-\alpha) hist_t + \alpha hist_{new}. \tag{6}$$

At each frame a threshold is applied on the pixel values in the region *R*, and then it is calculated the percentage of relevant pixels in each direction. In that way, bottom and upper pixels between the regions *R* and *Q* are considered to belong to *y* axes, and lateral pixels between the regions *R* and *Q* are considered to belong to *x* axes. To avoid a sudden variation of the window size, it is done a weighted mean between the current window size and the new scale. Thus, firstly, it is calculated the new scale *s* using the moments and a constant *k*

$$s = k\sqrt{M_{00}/(256wh)}. \tag{7}$$

Then, it is employed (8) and (9) to perform the weighted and to update the size and shape of the search window, where *t* and *t+1* are indexes that correspond to the current dimension and next dimension.

$$w_{t+1} = w_t (1 + \beta(s-1))(1 + 2\eta(perc_x - 0.5)). \tag{8}$$

$$h_{t+1} = h_t (1 + \beta(s-1))(1 + 2\eta(perc_y - 0.5)). \tag{9}$$

The variables $perc_y$ and $perc_x$ are the percentage of relevant pixels in each axes, and β and η are constants.

III. THE SYSTEM

To track targets in the presence of occlusions, we combined Kalman filter and Camshift. Fig. 2 illustrates the finite state machine (FSM) that controls the system.

The transition between states depends on the *A* value, where *A* is the percentage of relevant pixels inside of the search window (inside of the regions *Q* and *P*). To obtain this value, it is applied a threshold on the pixel values. The rate between the number of relevant pixels and the total number of pixels gives us the *A* value. According Fig. 2, if the system is in state 1 and *A* value is less than A_{13}, then the next state will be 3, and so on.

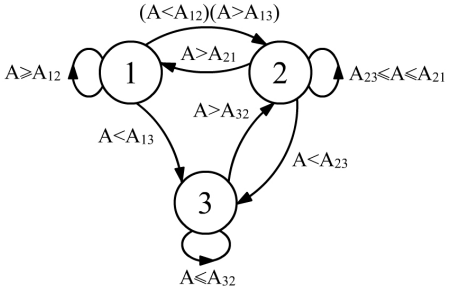

Figure 2. Finite state machine of the system.

Initially, the system starts in the first state. In this state, the tracker is guided by Kalman filter, whereas Camshift gives the next information to permit the tuning of Kalman filter. Moreover, the histogram and window size are updated at each frame. A similar procedure happens in the second state, but the histogram is not updated. Finally, in the third state, the tracker is guided by Kalman filter and both histogram and window size are not updated. When the FSM goes to third state this means that the system lost the target. So, in order to add a certain level of uncertainty, whenever this happens the window size is expanded by a factor K_{inc}. When the FSM passes from state 3 to 2, that is, when the tracker finds the target, the window size is reduced by a factor K_{dec}.

When the FSM is in the third state, it is considered that the observations are missing. Then, we use the previous states found by Kalman filter. This procedure was checked as the best choice according to [5].

IV. FINDING THE HUMAN FACE

Find human face in video sequences is an important procedure to identify people presence. The procedure proposed in this work to find the human face is based on the structure of the human body. Once the body is found by the system proposed in Section III, its human face is always sought in the first state of FSM. Then, another search window is used to find the face, whose initial dimensions are proportional to the dimensions of the search window of the system. The initial localization of this window is above of the vertical centroid and it is centered in the horizontal centroid.

Although this procedure gives us a good localization of the human face, it may also bring us some additional information, like background or body of the person, which may be undesirable. In order to improve this result, we apply a refinement technique of the human face position. This technique is based on work presented in [6], and it uses hue channel to segment skin. Two threshold values are employed on the hue to segment the image. Then, we apply one iteration of Camshift to find the new location and the new dimensions of the search window. In that way, we hope that a human face stays located in a more centered position on the window. However, if the zero order moment has a value below of a threshold, the window will be discarded.

This technique has the advantage that it is fast. However, it does not present good results for target tracking, especially in outdoor environment. It is not difficult to find objects whose response of the hue channel is similar to human skin. Nevertheless, the main reason to choose the hue to segment human skin is that this channel is almost invariant among the people [6][7].

V. EXPERIMENTS

In this section, we describe the carried out experiments and we show the achieved results. The parameter values in the experiments were the same used in [2]. The histograms of the HSV color space were divided in bins of size 32x32. However, in the back-projection step, we used only the 8 greatest bin values of each channel to become the system faster, and the other pixel values was set to zero. In the Kalman filter, we employ the matrix transition described in (10), where d_{t+1}, d_t and d_{t-1} are the next position, current position and past position, respectively,

$$d_{t+1} = 2d_t - d_{t-1}. \quad (10)$$

In this work, we employ an automatic version, without human interference in the system, that is, all procedure is accomplished automatically. In order to acquire initial information about position of the search window, we evaluated three methods: Ogale-Aloimonos optical flow (OA-optical flow) presented in [8], Lucas-Kanade optical flow (LK-optical flow) [9] and frame subtraction. All algorithms were applied on the gray scale of the images. One advantage of using optical flow is the possibility for choosing the "speed" of the object of interest. Despite of the subtraction of frames to be a simple and fast procedure, it is not robust. One drawback of these methods is that the object needs to be in movement, and the background and the camera need to be fixed. In other case, the speed difference between the object and the background must be relevant.

These methods are used between the current frame and the previous frame. Furthermore, it is used a morphological operator to remove noise in this process [10]. When it is found an object with similar structure to the object of interest, it is calculated the centroid of the found target. With the computed centroid, we use only the region near to centroid to obtain the probability distribution image for the two channels of the HSV space. Then, we obtain the location and color of the target, and we can use the proposed system.

In order to evaluate the capability of our approach for human tracking, we have tested our approach on 140 video sequences from Gatech [11]. The result obtained for each approach is shown in Table I, where the column “tracking” gives the percentage of video sequences that the system obtained success in the tracking task. The result obtained in [2], using OA-optical flow on the same set of videos, was added for comparison.

As we can see in Table I, our approach obtained a performance up to 62.85% with OA-optical flow. Moreover, the quality of tracking with this optical flow was better than the from other ones. The performance obtained by frame subtraction was as good as the obtained one by LK-optical flow, furthermore, it was the fastest algorithm among them. Fig. 3 shows some frames of the video sequences.

Comparing with the approach presented in [2], where Camshift guides the tracker, instead of Kalman filter, our approach obtained better results. Furthermore, in this method the most time only one iteration of Camshift was necessary to give information to Kalman filter. Therefore, this approach has a low computational cost if compared with approach proposed in [2].

In addition, we tested our system with OA-optical flow on another video sequence (Fig. 4). Different from Gatech videos, this video has target occlusion and there is rotation of the camera around of the vertical axes. These details become difficult for optical flow tracks the target. However, our approach has obtained success.

TABLE I. PERCENTAGE OF SUCCESS FOR EACH TECHNIQUE

Technique	Tracking (%)
OA-optical flow	62.85
LK-optical flow	44.28
Frame subtraction	42.85
Result in [2] using OA-optical flow	55.71

To find the faces of the persons in the videos, we have used the values 3 and 43 as lower and upper thresholds of the hue. These values were found experimentally in [6]. Fig. 5 and 6 show situations of true positive and false positive face detection, respectively. In many situations, the heads were detected with little presence of either background or body, as we can see in Fig. 5. We emphasize that this task is difficult in the used video sequences due to low resolution of the human faces.

VI. CONCLUSION

In this paper, we present a combination of Camshift and Kalman filter to automatically identify and track people walking and running in outdoor environment. In addition, we propose a fast method to find the faces of the persons in video sequences. We carried out a series of experiments and we have obtained satisfactory results. In future researches, we will study some more robust methods for human tracking and find human faces.

ACKNOWLEDGMENT

This work is partially supported by the Fundação de Amparo a Pesquisa do Espírito Santo – FAPES – Brasil (Grant 41936450/2008). P. M. Ciarelli thanks at PPGEE (Programa de Pós-Graduação da Engenharia Elétrica) of UFES.

REFERENCES

[1] C. Yam, M. S. Nixon, and J. N. Carter, "Performance analysis on new biometric gait motion model," IEEE Southwest Symposium on Image Analysis and Interpretation, vol. 0, p. 0031, 2002.

[2] P. M. Ciarelli, E. O. T. Salles, and E. Oliveira, "A modified supervised and unsupervised Camshift for human motion tracking outdoor," ISSNIP: Biosignals and Biorobotics Conference, p. 6, 2010.

[3] M. J. Swain and D. H. Ballard, "Color indexing," International Journal of Computer Vision, vol. 7, no. 1, pp. 11–32, 1991.

[4] M. Zhao, J. Bu, and C. Chen, "Robust background subtraction in HSV color space," Multimedia Systems and Applications V, vol. 4861, no. 1, pp. 325–332, 2002.

[5] A. H. Jazwinski, "Stochastic processes and filtering theory". New York: Academic Press, 1970.

[6] F. Dadgostar, A. Sarrafzadeh, "An adaptive real-time skin detector based on hue thresholding: a comparison on two motion tracking methods," Pattern Recognition, vol. 27, pp. 1342–1352, 2006.

[7] G. R. Bradski, "Computer vision face tracking for use in a perceptual user interface," 1998.

[8] A. S. Ogale and Y. Aloimonos, "A roadmap to the integration of early visual modules," Int. J. Comput. Vision, vol. 72, no. 1, pp. 9–25, 2007.

[9] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," IJCAI81, p 674–679, 1981.

[10] R. C. Gonzalez and R. E. Woods, "Digital image processing," second edition. New Jersey: Prentice Hall, 2002.

[11] "Georgia institute of technolgy." [Online]. Available: ftp://ftp.cc.gatech.edu/pub/gvu/cpl/walkers/subjects/.

Figure 3. Some frames of the video sequences. Black windows are the search windows of the targets and white windows are search windows of faces.

Figure 4. Frames 63, 91, 168 and 336 of the video sequence with partial occlusion and rotation of camera.

Figure 5. Some true positive faces.

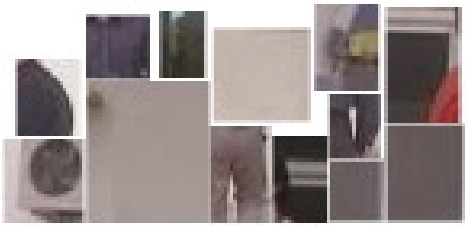

Figure 6. Some false positive faces.