ATLANTIS PRESS
Advances in Computer Science Research, volume 63
International Conference on Artificial Intelligence and Engineering Applications (AIEA 2016)

Crowd monitoring based on Face Orientation Estimation

Xiaodong Wang [a], Hong Xiao [b], Rongxiao Guo [c], Jie Cui [d]
AFEU University, Xi'an, 710077, China.
[a]wangyd1974@sina.com, [b]siaohong@126.com, [c]Lzstella26@163.com, [d]watermelon_cj@163.com

Abstract. Aim to the different needs, crowd monitoring and control is regarded to the key of modern surveillance. But, existing methods are still not good enough to be used on the large-scale spot, for the reasons of effect and efficiency. In fact, witnesses, who stay on the spot, practically act as "smart" sensors to the incidents. Thus, to detect their responds to events are effective to surveillance. Among all kinds of human actions, face orientation is easy to be detected; meanwhile it also implies plenty of immediacy with respect to the event. This paper presents a method of crowd monitoring based on face orientation estimation. According to preliminary experiment, it is proved valid and feasible.

Keywords: Crowd Monitoring; Event; FOE; Overlap; Decay.

1. Introduction

To the different needs, such as security, economy, health, and so on, crowd monitoring and control are essential. Unfortunately, valid solutions are still lack. The reasons, why crowd monitoring is so difficult, mainly include varied objects, open space, complex relevant data analysis, and even ethical consideration. Moreover, the status of the spot under the monitoring usually keeps changing, so that the real-time monitoring and control are more hardly implemented. Some new technologies and idea ought to be imported to push it forward.

Based on some investigation and relevant research in [1], FOE (Face Orientation Estimation) is deemed to enable to solve the problem mention above, because this technology is characteristic of the simple principle, mature algorithm and especially the advantage of surveillance without privacy invasion. So, in this paper, the method of crowd monitoring and control based on FOE is presented.

2. State of the art in FOE

FOE (or head pose estimation) is an important topic in computer vision and pattern recognition. It is an important area of research in human computer interaction (HCI). There are many researches in the area of estimation with monocular vision [1]. Methods for head pose estimation can be classified into two main categories: model-based and face property-based. In summary, the property-based methods are simpler, but less accurate; many require a large number of training face image under different orientations. Model-based methods usually start with feature detection, followed by matching 2D/3D corresponding features and determining face pose using the matched features. Among all facial features, the most commonly used eyes, nose, mouth [2-3].

3. The discussion of crowd monitoring

How to automatically find out incident over wide and open space is a problem. Existing methods depend on the complicated situation data analysis and calculation (voice, heat, movement, etc.), but little of them work well. To some degree, those data cannot really reflect the event at all or the information implied in the data is hardly calculated out by present machine and algorithm. Why not to watch the humans over the spot whose responses are directly relevant the event around them?

3.1 Learning from humanity.

An individual may find event by his eyes, ears, and nose. Combined these information with his brain reasoning and experience, he can detect most of incidents around him. In the other words, the ones on the spot are equal to "smart" sensors. To find out what and where an event happens, it is

another way to monitoring the attendee's respond besides direct evnet monitoring. This idea rounds the problems of huge data collecting and the accompanying complicated analyzing.

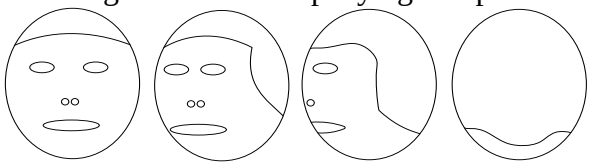

Figure 1 Illustration of face orientation

3.2 Extract the key information with respect to the events from human actions.

The action of humanity is complicated and diverse, from facial expressions to body movement. Practically, some tiny action is not suitable to be caught, while some imply too little information. To the best of our knowledge, the face orientation is very useful to the monitoring, because human usually prefer to face to (positive ones, such as interesting, strange and so on) or deviation from (negative ones, such as dangerous, terrible, and so on) the direction of incidents happen. Furthermore, having clarified the event, the attendees' subsequent actions are very coincident. There exist two cases, one is to turn his face to it in order to collect more information about the event, and the other is to turn to the inverse direction to escape from it. Figure 1 shows several kind of status of face orientation.

3.3 Face orientation monitoring to detect event

An event usually is not found by just one but several ones at same time (especially over wide and open space). All of them are apt to respond it. Thus, if we monitor most of the human among the spot, and compute their faces orientation overlap, we will find out whether and where event happen. The illustration is shown in figure 2. This kind of monitoring works without any privacy invasion. In the figure 2, three ones' face to the event together. Their sights form an intersection, which implies what the event is and where it happens.

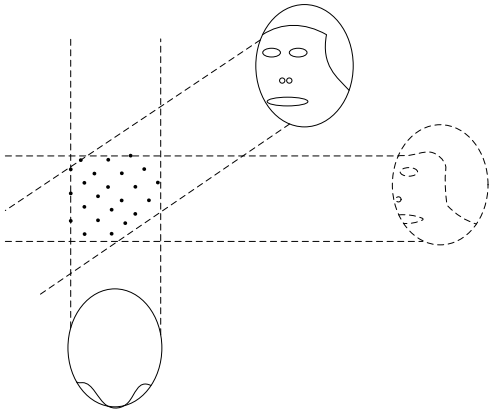

Figure 2. Sights form an intersection

Obviously, as long as to find out the intersection, we may detect incidents. Current available technologies (namely FOE) completely afford to the detection and corresponding calculations, so it is feasible.

4. Crowd monitoring based on FOE

The general framework of our method is illustrated in the Figure 3, which includes six steps. It mainly analyzes every individual among crowd and then detects their focus according to FOE to calculate out the position of event with respect to the focus.

4.1 Capture living image at Intervals

In order to keep continuous monitoring to the spot, images need to be captured at intervals, denoted to t. The following steps will analyze every image in turn, and then compare the result to find out the status changing with respect to the incidents. Herein, the short t means high calculating cost; on the

contrary long one induces information lost. So the interval t is set through prior computing to optimal.

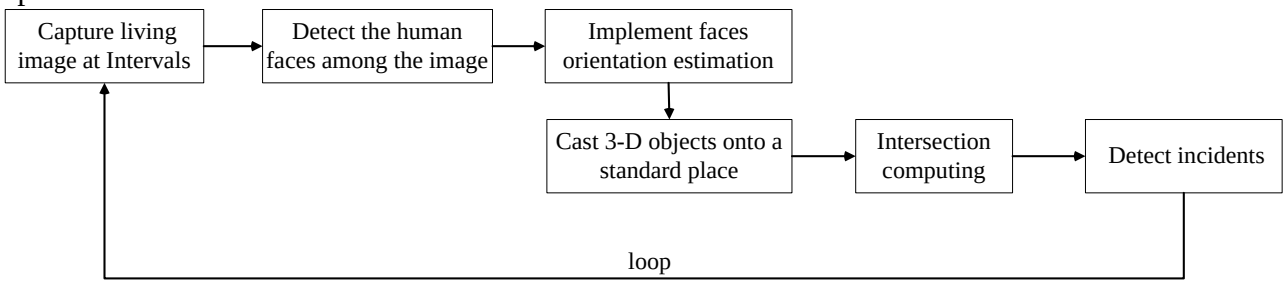

Figure 3. General framework of our method

4.2 Detect the human faces among the image

To an image i, we regard everyone in it as a "smart sensor" to incident, so these objects ought to be picked out from i firstly. There exist a handful of methods to detect human face among image, namely face tracking, which are mature and widely used in different field [4]. By these methods, we can detect human faces among i.

4.3 Faces Orientation estimation

In this paper, we do not plan to implement the whole face recognition or eye tracking, but just face orientation estimation. Because the latter is simpler than the former, and the calculate result is enough to support to the following steps. Once the orientation is estimated, one's attention area may be found out by the FOE and then will be denote on the image i, shown on left of figure 4 (several these areas may form a focus like the right one). In the figure, the black point e denotes to the concern or focus, l to the distance between i and person. Considering the object size, the sight is given a certain width, so that it forms an attention area.

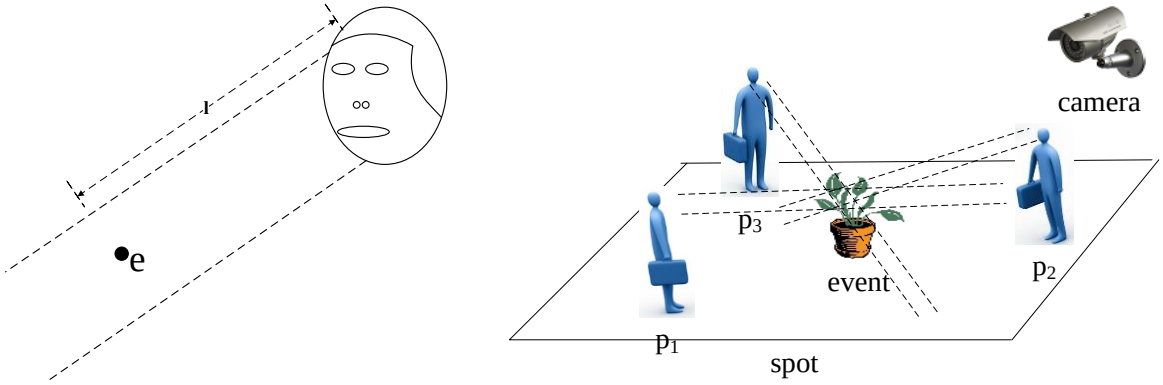

Figure 4. The illustration of Attention calculation

The attention area which covers all of objects is accorded with the fact orientation. Herein, the attention degree is inversely proportional to the distance l. We calculate the attention of point e with formula (1):

$$\alpha_p = \theta / l \tag{1}$$

Herein, α_p is the attention degree of person p, θ is the constant. Obviously, the attention value decays with the distance between point and observer. Estimate all of the faces over the spot and draw their attention areas.

4.4 Cast 3-D objects onto a standard place

Obviously, in figure 5 the attention areas with angle are not very suitable to the intersection calculation. Herein, in order to simplify this calculation, we prefer to cast them onto a standard place. Because we may attain the angles from the camera initial parameters (which include the view angle of the equipment), these angles are looked as to known. By these parameters, the attention areas are cast to the standard place shown in figure 5.

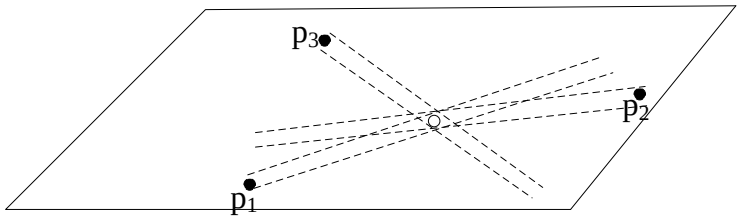

Figure 5. The attention cast on place

4.5 Intersection computing

In figure 8, we may find the intersection easily. In order to exactly attain the intersection, the following sub-steps will be carried out. (a) computing attention degree value, with respect to a person, of every point on the place. With the formula (1), the attention degree of every point on the place may be gotten. Herein, the attention degree is just with respect to one person. By this way, all of attendee's attention degrees with respect to every point on the place are calculated out in turn. (b) sum up all persons' attention degree together. In this step all person's degrees are sum up with the formula (2):

$$A = \sum_{i \in P} \alpha_i \quad (2)$$

Herein, A is total attention degree with a point on the place, P is the attendee set, $P = \{p_1, p_2, \cdots\}$, p denote to an attendee. Finally, the distribution map of attention degree is drawn out.

4.6 4.6 Detect incidents

With help of the distribution map, we may find the interaction mentioned above which mean where the incident happens. There exist two cases to analyze the incident. 1) based on the density: In this case, the higher the attention degree, the more possible it is an incident area. 2) based on the changing: In this case, the different images analyses are connected. It deems the place, the attention degree of which sharply change, deduce to an incident area. The two cases may be respectively used to positive and negative event monitoring.

5. Experiments

In this part, some experiments are designed to test the effect of the method.

5.1 Experiment design

The system is input no video but image as object so as to simplify the process. The images are collected from Internet by search engine, with the key words "gather" or "assemble". The images, which pass the review and regarded to suitable to the test (3 conditions should be satisfied as follow: a. every image only has one insight focus; b. incident may be reflect by the attendee; c. the image is suitable to implement EOF), are stored into testing set. Face++ (http://www.faceplusplus.com/) is adopted to implement EOF and herein, we assume that the 3D space parameters, with respect to the image, are known, as they may be attained when the cameras are assigned. Herein, we will not discuss the 3D parameters in detail. Thus the calculation mainly focuses on the calculation of distribution map of attention degree.

The experiments will test the validity and time consume of the method.

5.2 Experiment results

8 images are extracted from the image set at random. All of the image will pass the whole processing and the results are synthesized to the analysis report.

The method tracks the faces on it, and then sum up all of insights to calculate out the total attention degree among 3D space. The place which has the highest attention degree is deemed to the event location. In order to quantize the accuracy, we compare the result between machine and human with the formula (3):

$$\varepsilon = \Delta l / L \quad (3)$$

Herein, ε is Deviating Coefficient (DC), Δl is the distance which machine result deviates from the human's, L the total length of the image. Table 1. Shows the accuracy of our method.

Table 1. The accuracy test

Image	1	2	3	4	5	6	7	8	average
DC	0.01	0.12	0.22	0.01	0.03	0.05	0.11	0.07	0.08

The time consume is very important to monitoring system. If not rapid enough, it means the method is useless. The time consume test result is shown in table 2. Though it is heavily relevant to the image content, the average time consume is still acceptable. The main time-consuming processes are EOF and 3D space mapping calculation.

Table 2. The time consume test

Image	1	2	3	4	5	6	7	8	average
TC(ms)	220	308	444	219	657	322	198	224	324

In general, the method is proved to be valid and feasible, but some defects also should be noticed to be improved in following days. 1) As some objects cannot be recognized, some negative score are contributed to the calculation by them. The FOE needs to be adjusted for this method. 2) The information from neighbor frames nearby, are not considered, which maybe improve the accuracy. 3) The time consume is relative high, the algorithm ought to optimized.

6. Conclusion

In order to solve the problem of crowd monitoring on large-scale spot, this paper takes the man on the spot as "smart sensors" to the incidents. With the help of EOF, their insights are collected to form common focus, which is regarded as the location of the incidents. In order to improve the effect and accuracy, some other technologies, include decaying estimating, 3D image and tracking, are involved in. The results of experiments illustrate it is feasible and effective. But, this just means a start, and some other more works need to be carried out, such as recognition, 3D space mapping and relevancy consideration between different intervals.

Acknowledgements

This paper is supported by the basic research programs of science and technology commission foundation of Shaanxi province (NO. 2012JM8013).

References

[1] Qiang Ji. 3D Face pose estimation and tracking from a monocular camera. Image Vision Computing. 2002. No. 20. pp. 499 – 511.

[2] Gee A., Cipolla R. Fast visual tracking by temporal consensus//Image and Vision Computing. 1996.– No. 14. pp. 105-114.

[3] Horprasert A.T., Yacoob Y., Davis L.S. Computing 3D head orientation from a monocular image sequence. Proceedings of SPIE 25th, AIPR Workshop: Emerging Applications of Computer Vision. 1996. No. 2962. pp. 244 – 252.

[4] Harihara krishnan K, Schonfeld D, Raffy P, et al. video tracking using block matching. In:IEEE International Conference on Image Processing, 2003, 945-948.