

ATTENTION-BASED DENSE CROWDS ANALYSIS

Matei Mancas

University of Mons, IT research center/TCTS Lab
20, Place du Parc, 7000 Mons, Belgium

ABSTRACT

The use of algorithms which model part of the human attention can bring interesting results in the video analysis of difficult scenarios like dense crowds. A rarity-based attention model is able to provide in real-time areas in video frames where motion behavior is surprising compared to the rest of the motion in the same frame. This algorithm is also resistant to camera shake or translation and points out abnormal activities which can be used in surveillance but also to analyze and even foster social interaction.

1. VIDEO PROCESSING IN DENSE CROWDS

Video processing for dense crowds is a field of computer vision which has specific properties. It is for example virtually impossible to obtain individual object tracking or it is difficult to acquire databases of specific events. If the main applications of crowds monitoring remain in video surveillance, applications around social signal emerge.

A first category of papers in the field is related to crowd properties analysis, and a second one to abnormal event detection. Within the first category, a lot of papers estimate crowd density using textures, edges, or global cues [1, 2] or using optical flow [3] to detect stationary crowds. Some people counting in crowds results were also achieved [4].

In the second category, the aim is to detect and if possible to classify abnormal events in crowds. Most of the time, normal behaviors are modelled and deviations from those models are considered abnormal. In [5] authors use HMM and principal component analysis. In [6] an interesting approach uses lagrangian particle dynamics for the detection of flow instabilities and the method seems to be efficient for dense crowds. [7] uses optical flows to detect when abnormal events occur without necessarily pointing the precise region of interest into the frames.

Our approach is a real-time contribution to abnormal event detection and uses the notion of computational attention which quantifies motion saliency. It is possible to precisely locate the area into the crowd where abnormal or surprising events occur. In sections 2 and 3 computational attention is defined and discussed. Section 4 details the proposed method while section 5 presents some results. Section 6 concludes with several applications and the possibility to use the method for social interaction analysis in crowds.

2. COMPUTATIONAL ATTENTION

The aim of computational attention is to automatically predict human attention on multimodal data such as sounds, images, video sequences, etc... The term *attention* refers to the whole attentional process that allows one to focus on some stimuli at the expense of others. Human attention mainly consists of two processes: a bottom-up and a top-down one. Bottom-up attention uses low-level signal features to find the most salient or outstanding objects. Top-down attention uses a priori knowledge about the scene or task-oriented knowledge in order to modify (inhibit or enhance) the bottom-up saliency. While numerous models were provided for attention on still images, videos have been less investigated. Nevertheless, some authors generalized their models to videos ([8, 9] present a more detailed review of saliency algorithms in videos). Most of these methods are mainly applied on more classical mono- or multi-user scenarios and not on dense crowd scenarios. The presented approach uses the instantaneous spatial context: it compares a given motion behavior to the rest of the motion within the same frame.

3. MOTION DETECTION VS. PERCEPTION

Motion is a very important feature in human perception and it highly attracts human attention. But as any other feature and as already stated in [8] it does not attract attention by itself: bright and dark, locally contrasted areas or not, red or blue can equally attract human attention depending on their context. In the same way, a high amount of motion can be as interesting as little motion depending on the context. The main cue which involves bottom-up attention is the rarity and contrast of a feature in a given context. The feature considered in this paper is motion speed. This feature is defined in the method presentation in the next section.

A low-computational-cost quantification of rarity was achieved [8] referring to the notion of self-information. Let us note m_i a message containing an amount of information. This message is part of a message set M . The bottom-up attention attracted by m_i is quantified by its self-information $I(m_i)$ which will be called here *saliency index*:

$$I(m_i) = -\log(p(m_i)) \quad (1)$$

where $p(m_i)$ is the occurrence likelihood of the message m_i within the message set M . $p(m_i)$ is estimated as a combination of the global rarity of m_i within M and its global contrast compared to the other messages. The current implementation only uses the global rarity, thus $p(m_i)$ is:

$$p(m_i) = \frac{H(m_i)}{\text{Card}(M)} \quad (2)$$

where $H(m_i)$ is the value of the histogram H for message m_i and $\text{Card}(M)$ the cardinality of M . The M set quantification provides the sensibility: a smaller quantification value will let messages close to each others to be seen as the same. The saliency index (or motion attention index, $I(m_i)$) operates at three levels corresponding to three different time scales: up to 1s (instantaneous motion attention), from 1s to 3s (short-term motion attention), more than 3s (long-term motion attention). In this paper, as a first approach, only the instantaneous motion attention is used. Let us consider a crowd with interacting people. Motion features (here motion speed) characterizing each moving scene area are compared at each instant. Salient motion behavior (e.g., one person speed very different from the others) immediately pops-out and attracts attention. This refers to pre-attentive human processes, usually faster than 200 milliseconds. In this approach, motion saliency detection at instantaneous level is computed over time intervals of 200ms – 1s.

4. METHOD

The presented method is based on a two-step analysis of the perceived motion speed of different areas within the crowd. In a first step, the crowd movements are segmented according to their motion speed and spatial density. In a second step, a rarity-based bottom-up attention approach is used to highlight the most salient motion areas in the crowd.

4.1. Crowd motion grouping

This technique only considers moving objects, thus completely static areas in crowds could not be detected as salient even if they are rare. Nevertheless low-, mid- and high-speed motion can be well detected by using the standard approach of frame-differencing. As no scene background model is available and it is impossible to model one during crowd evolution, background subtraction cannot be achieved. The only motion feature used here is the motion speed and no other cues as the motion directions were taken into account. Motion speed is not easy to define in real-world scenarios and especially in crowds, thus, the speed is here extrapolated from the quantity of motion in the scene which can be measured by the area occupied by the movements reported to the object area. As individual objects in crowds are impossible to segment but they should have similar sizes, the perceived motion speed can be quantified by the local quantity of motion or the local movement area



Fig. 1. Left: original frame; Right: frame-differencing history.

size. As it can be seen in Figure 1, larger the white area is in the right image, higher is the motion speed of the object. To extract blobs by size from a binary image, mathematical morphology is a very efficient technique. The video frames are reduced to a 320 pixels wide resolution to keep the same parameters for all the videos. The size of the white areas (implying a different motion speed) will be divided into three classes: low-, mid- and high-speed of motion as it can be seen in Figure 2.

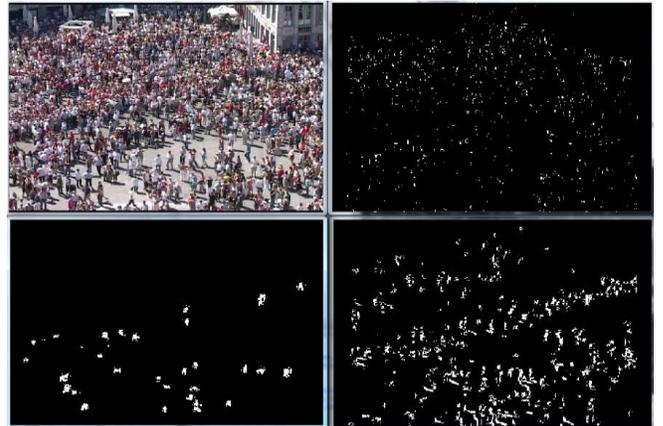


Fig. 2. Top-Left: original frame; Top-Right: small blobs (low-speed); Bottom-Right: medium blobs (medium-speed), Bottom-Left: large blobs (high-speed).

To achieve this goal, morphological opening with growing structural element sizes are used to select blobs of a given size (Figure 2). A first morphological opening using a small structuring element separates low speed motion from the rest while a second opening operation with a larger structuring element is able to separate the mean from the high speed areas. Once the speed maps extracted, a second step provides information about blobs' neighborhood.

A low-pass filtering with a large square kernel (51 in this implementation) is applied on each of the motion speed maps in order to quantify blobs density. The filtering result is higher if the blob neighborhood is crowdie with other blobs and smaller if the blob is alone. A threshold on the filtering results provides for each of the three motion speed maps two sub-maps showing blobs which are isolated (low-pass filtering smaller than the threshold) or grouped (low-pass filtering result higher than the threshold). An example of this decomposition can be seen in Figure 3 by comparing the first column (isolated blobs) to the second one (dense blobs) for each one of the three motion speed maps (rows).

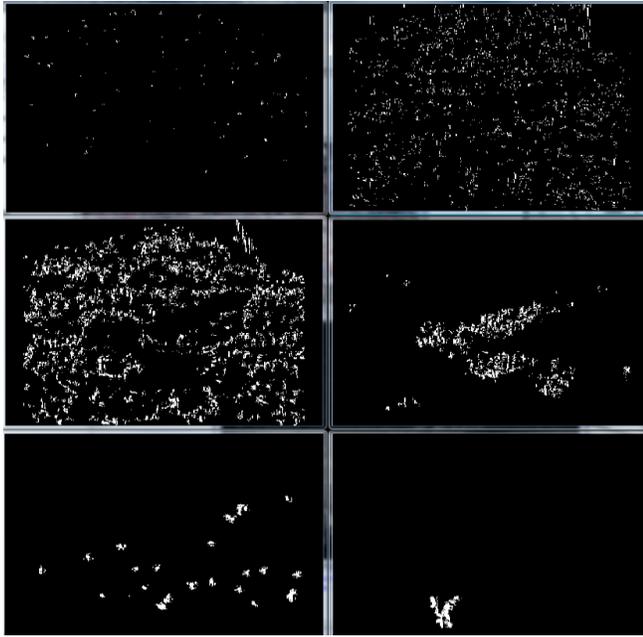


Fig. 3. First row: small speed areas having low density (left) and high density (right); Second row: mean speed areas having low density (left) and high density (right); Third row: high speed areas having low density (left) and high density (right).

4.2. Motion bottom-up attention

The detected motion (Figure 1, right image) of the original frame (Figure 1, left image) is thus grouped into six maps which provide information about the moving crowd areas' speed and density (Figure 3) for this frame. The bottom-up rarity approach described in section 3 is here used to decide which one of those six maps has the lowest occurrence and is by consequence the most surprising or salient. A simple and computationally efficient way to deal with that issue is to count the number of blobs in each map and to divide it by the total number of blobs. Then it is possible to apply Eq.1 and Eq.2 from section 3 to obtain the saliency index of each of the six motion speed maps. Finally, the six maps saliency indexes are superimposed to get a unique bottom-up map as it can be seen on the right image of Figure 4. Clear areas are rare or surprising compared to the rest of the frame, thus they have a higher saliency index. Darker areas are less salient. Black areas are the "no motion" regions which are not taken into account in this implementation.



Fig. 4. Left image: original frame; Right image: global result: clear areas have a higher saliency index than the darker ones.

5. RESULTS

This technique is based on crowd motion grouping and reliable results to this first step are needed to obtain comprehensive global results. The parameters of the current implementation are tuned for dense crowds (as those which can be seen on figures 1, 4, 5 or 6 on the left images) and will need parameter adaptation to work on less dense crowds where people have a much larger size.

A real-time implementation was achieved by using the EyesWeb XMI platform (www.eyesweb.org) and tests were made with crowds in various scenarios (religious and cultural manifestations, concerts and football fans).

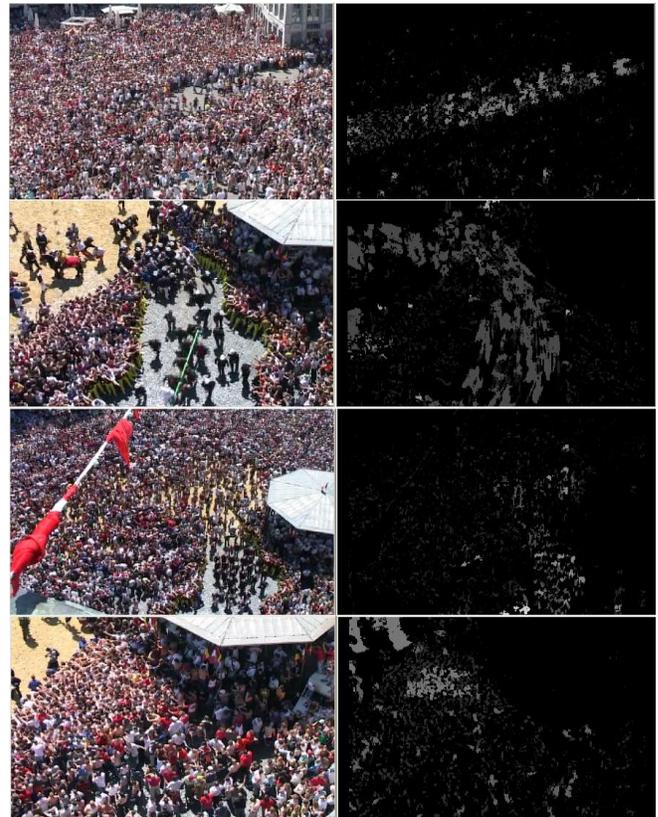


Fig. 5. Left column images: original frames; Right column images: global result of saliency index maps.

The use of higher spatial resolution needs higher computational load without benefits on results precision. Even if very important features as motion direction are not taken into account here, the first results are subjectively close to human motion perception. Motion grouping clearly highlight groups of people with similar behavior and the basic rarity quantification provides convincing results as shown in Figure 5. The original frames are on the left column and the corresponding bottom-up saliency index on the right column. The motion flows which are locally packed are highlighted, but it is possible also to see more isolated people with higher speeds than their neighborhoods inside or outside those flows (rows 1, 2 and 3).



Fig. 6. Left image: original frame; Right image: global results: camera translation has very little influence on the algorithm results.

For row 4 (Figure 5), a crowd sub-group is highlighted as it has a different behavior compared to the majority. An interesting point is that naturally, low speed motion is found in majority in most of the cases which leads to higher saliency indexes to areas with higher speeds but also to areas with dense groups even if they have a medium speed. Another interesting point is that as the majority of the motion has a very low saliency index, small camera movements as vibrations or translations have very few effects on the saliency index map. Figure 6 shows a scene with a horizontal movement of the camera from right to left (see the left column images). Almost no noise due to this global translation affects the saliency map which focuses on some synchronized individuals in the front.

6. DISCUSSION AND CONCLUSION

Behavior analysis of dense crowds has a lot of applications ranging from classical ones to other innovative applications. Surveillance based on this approach could point out the hot region of interest within the crowd to a human operator. Crowd smart event summarization and coding could provide more importance to the abnormal motion areas.

Another application is more trans-domain and aims in studying the social behavior of crowds: it could be possible to follow abnormal events in the crowd and see if their evolution leads more to contamination to the rest of the crowd (progressive saliency index decrease) or to extinction (saliency index disappears after a short period of enhancement). Moreover, it should also be very interesting to provide the crowd with a feedback of the system: in that way some groups of people will be pointed out and provided either with positive or negative feedbacks. The social

behavior of the crowd will certainly be highly influenced by the system feedback and the analysis will show if the tendency is to global synchronization or on the contrary to emerging novel behaviors. If the feedback has an artistic value, by influencing this feedback, the crowd could achieve common public art... A real-time bottom-up saliency index which models part of the human attention was demonstrated on dense crowd videos. In a first step, this method provides motion decomposition according to its speed and density and in a second step, the rarity of the different components of this decomposition is used to highlight the abnormal or surprising behaviors. The results are surprisingly close to a human observer perception despite the fact that only speed-related features and bottom-up attention were used. It is possible for example to detect sub-groups of people having abnormal activity into the crowd and the method is robust to small camera movements.

In addition to the use of other motion features like directions, a very useful improvement of this method is to add top-down information by modeling usual motion in order to inhibit part of the bottom-up attention. Additional evaluation of the algorithm is also required as for example the use of predefined ground truth in synthetic data [6].

11. ACKNOWLEDGEMENTS

The research program “Numediart” (www.numediart.org) is funded by the Walloon region, Belgium.

12. REFERENCES

- [1] A. Marana, S. Velastin, L. Costa, and R. Lotufo. “Estimation of crowd density using image processing,” *Image Processing for Security Applications*, IEE Colloquium, pages 11/1–11/8, 1997.
- [2] R. Ma, L. Li, W. Huang, and Q. Tian. “On pixel count based crowd density estimation for visual surveillance,” *Cybernetics and Intelligent Systems*, 2004 IEEE Conference, vol. 1:170–173, 2004.
- [3] B. Boghossian and S. Velastin. “Motion-based machine vision techniques for the management of large crowds,” *Electronics, Circuits and Systems*, vol. 2:961–964, 1999.
- [4] S.-F. Lin, J.-Y. Chen, and H.-X. Chao. “Estimation of number of people in crowded scenes using perspective transformation,” *Systems, Man and Cybernetics*, IEEE Transactions, 31(6):645–654, 2001.
- [5] E. Andrade, S. Blunsden, and R. Fisher. “Hidden markov models for optical flow analysis in crowds,” *ICPR 2006.*, vol. 1:460–463, 2006.
- [6] S. Ali and M. Shah. “A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis,” *CVPR ’07.*, pages 1–6, 2007.
- [7] N. Ihaddadene, and C. Djeraba. “Real-time crowd motion analysis,” *ICPR 2008*, pages 1-4, 2008.
- [8] M. Mancas, D. Glowinski, G. Volpe, A. Camurri, P. Breteche, J. Demeyer, T. Ravet, P. Coletta. “Real-Time Motion Attention and Expressive Gesture Interfaces,” *Journal On Multimodal User Interfaces (JMUI)*, Springer Berlin/Heidelberg, 2009.
- [9] M. Mancas. “Relative influence of bottom-up and top-down attention,” *Attention in Cognitive Systems*, LNCS, Volume 5395/2009:pp. 212–226, 2009.