

Dense crowd analysis through bottom-up and top-down attention

Matei Mancas¹, Bernard Gosselin¹

¹ University of Mons, FPMs/IT Research Center/TCTS Lab
20, Place du Parc, 7000, Mons, Belgium
Matei.Mancas@umons.ac.be

Abstract. Video analysis of difficult scenarios like dense crowds can highly benefit from the use of algorithms which model part of human attention. Interesting motion which is new or surprising can be computed on large groups of people based on a two step approach. A bottom-up attention model built upon motion rarity compared to the rest of the motion in the same frame and a top-down approach which inhibits regions from the image which have a too repetitive behavior. This algorithm points out abnormal activities which can be used in surveillance but also to analyze and even foster social interaction.

Keywords: Visual attention, crowd, motion behavior, people, video, saliency.

1 Dense crowd analysis

Video processing for dense crowds is a field of computer vision which has specific properties as it is impossible to obtain individual object tracking or it is difficult to acquire databases of specific events. The main applications of crowd monitoring are in video surveillance but 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, global cues [1, 2] or optical flow [3] to detect stationary crowds. Some results in crowds counting 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. Saliency-based methods [8] on motion signal using both a bottom-up and top-down information were also presented but not on crowds and with a goal-oriented top-down influence.

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

2 From human attention to attentive computers

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, time-evolving two-dimensional signals such as videos have been less investigated. Nevertheless, some authors generalized their models to videos (for a detailed review, see [9, 10]). 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 both bottom-up and top-down information to model attention. The bottom-up approach uses the instantaneous spatial context: it compares a given motion behavior to the rest of the motion within the same frame. The top-down approach builds models of regions within the scene where some of the motion features are very repetitive. Those regions will thus inhibit the motion having the same features as the model in the same region.

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 [9] 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.

A low-computational-cost quantification of rarity was achieved [9] 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 1 second (instantaneous motion attention), from 1 to 3 seconds (short-term motion attention), more than 3 seconds (long-term motion attention). In this paper, the bottom-up instantaneous motion and the top-down long-term motion attention are used.

3 Model implementation

The presented implementation is based on a two-step analysis of the perceived motion speed of different areas within the crowd. In a first step bottom-up instantaneous attention is applied on the speed feature. In a second step, repetitive behaviors (in terms of speed) located in some scene areas are used to build models which are able to inhibit part of the motion in those areas if it is not novel. As individual tracking is not an option in crowd analysis, a motion grouping pre-processing is needed: the crowd movements are segmented according to their motion speed and spatial density.

3.1 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-, medium- and high-speed motion can be 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 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. Individual objects segmentation in crowds is an open challenge. Nevertheless, in crowds, individuals should have similar sizes, the perceived motion speed can thus be quantified by the local quantity of motion or the local movement area size. Figure 1 shows that larger the white area is in the right image, higher is the motion speed of the object.



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

In order to extract blobs by size from a binary image, mathematical morphology is an efficient technique. The size of the white areas (implying a different motion speed) will be divided into three classes: low-, medium- and high-speed of motion (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).

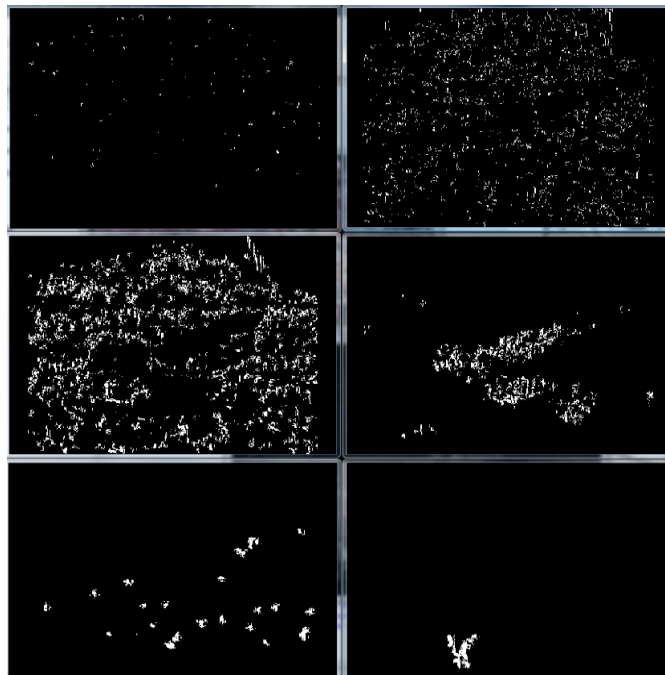


Fig. 3. First row: small speed areas having low density (left) and high density (right); Second row: medium speed areas having low density (left) and high density (right); Third row: high speed areas having low density (left) and high density (right).

To do this, morphological opening with growing structuring 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 with a larger structuring element separates the medium from the high speed areas.

Once the speed maps extracted, a second step provides information about blobs' neighborhood (Figure 3). 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 crowded 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).

3.2 Motion bottom-up attention

Let us consider a collective context, e.g., 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 area speed very rare compared the others) immediately pops-out and attracts attention. This refers to pre-attentive human processes, usually faster than 200 milliseconds [11].

To compare the scene areas' speed, 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).

The bottom-up rarity approach described in section 3 is here used to decide which one of those six maps has the rarest 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 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.

3.3 Motion top-down attention

Long-term memory (LTM) [12] component of the model processes the saliency index in a time interval from several seconds to much longer periods (related to the application time scale). The output is a modification of the instantaneous attention indexes in such interval according to their considered recurrence. While bottom-up attention mostly enhances rare behavior in space and it tends to suppress spatial repetitions, the top-down approach used here models the repetition suppression in time [13]. The repetitive behavior will thus be inhibited.

A model was already developed to compute the rarity for the direction and speed of participants observed in different regions of the space [14]. In the current implementation, only the speed feature is available, thus six long-term velocity models corresponding to the six maps obtained after the motion grouping step (Figure 3) are built. If the same motion velocity occurs many times in the same area of the scene (threshold parameter) this area is marked as having a usual velocity. If the same velocity occurs again in that area, it will be inhibited (usual motion) while if this feature occurs in a region where it is not usual, it will be highlighted (novel motion).

Figure 5 shows an image and five out of the six models (the first one is not shown because it contains very few information).

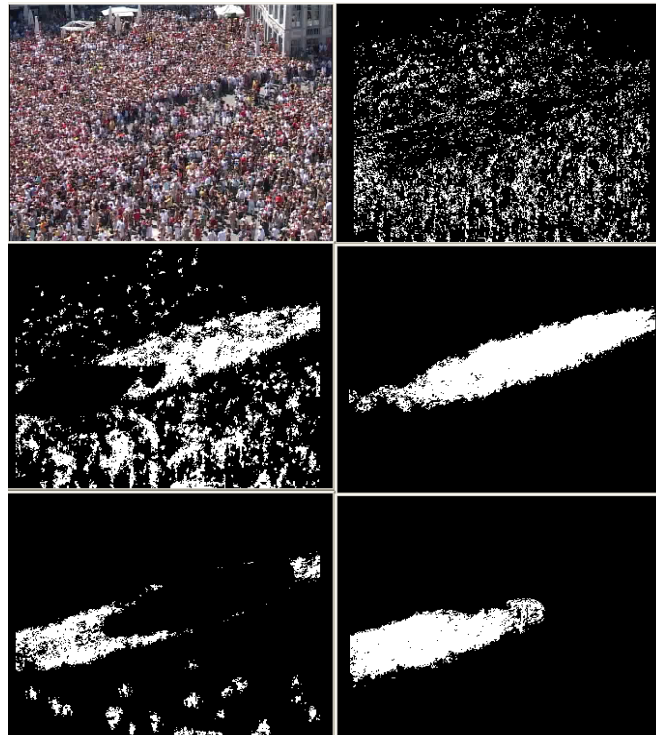


Fig. 5. Top-left: original frame; Top-down models of: on top-right: low velocity/high density areas, on middle-left: medium velocity/low density areas, on middle-right: medium velocity/high density areas, on bottom-left: high velocity/low density areas, on bottom-right: high velocity/high density areas.

The usual crowd movements in the middle of the image are visible on four models out of six. Those models will be subtracted from the corresponding motion maps from Figure 3 and only areas present in Figure 3 and not present in Figure 5 will be kept while the others are eliminated (inhibited).

During the first seconds, while the top-down models are not yet built, the top-down and the bottom-up maps are exactly the same. After some seconds, the model grows and if there are repetitive movements in a given location in the image, the top-down and bottom-up attention maps will be different.

Several parameters are important in the top-down models computation and they have to be adapted to the crowd context. The first parameter is the repetitiveness of the model: how repetitive a movement has to be in a scene area to be considered as being part of the model? As in crowds there is a lot of motion, the models should be built quite fast because human attention will be rarely available for a long time in a single place. A second important parameter which is an improvement compared to [14] is the speed used in the forgetting process. The models should be able to evolve in time and that is why a time constant is used to forget the older configurations. The forgetting speed is quite high because as there are many movements in a crowd, it is very difficult to remember motion precisely during a long time period. Finally a third important parameter is the maximum area in the scene which can be filled by movement before beginning to forget at a very high speed. This parameter is very interesting for example in the case where the camera has slight movements: the scene is suddenly filled with motion and this will lead a fast forget or “reset” of the model: if the camera moves, only bottom-up information is taken into account. It is important to notice that the forgetting process is not triggered in our implementation by a time limitation, but by a spatial limitation: if the model covers a too large part of the image, the forgetting process starts, otherwise nothing is forgotten.

4 Results

4.1 Bottom-up instantaneous motion attention

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).

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 used in the current implementation provide convincing results as shown in Figure 6. The original frames are on the left column and the corresponding bottom-up saliency index on the right column.

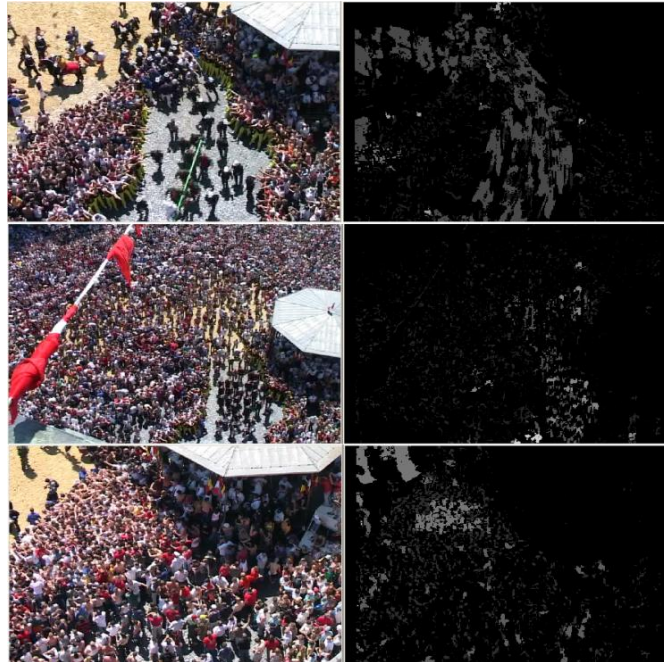


Fig. 6. Left column: original frames; Right column: result of saliency index maps.



Fig. 7. Left: original frames with global translation; Right: global translation results: camera translation has little influence.

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.

For row 3 (Figure 6), 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 7 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.

4.3 Top-down long-term motion attention

There are three possible influences of top-down attention on the bottom-up maps. If there is no special repetitive area in the scene, the top-down influence on the bottom-up attention map will be minimal.

A second top-down behavior is that if some areas are detected containing repetitive motion with the same velocity, part of the bottom-up attention map is inhibited. This is the case of Figure 8, where during the first seconds (when top-down models are not yet built) the bottom-up and top-down attention maps are the same (Figure 8, the top-right image). After some seconds, the models are built (here the medium velocity/high density areas model is shown in the bottom-left image).



Fig. 8. Top-left: original frame, Top-right: bottom-up attention map (and also top-down attention map during the first viewing seconds), Bottom-left: the medium speed/high to-down model, Bottom-right: top-down attention map after inhibition by the models.

In the final attention map after top-down modulation (Figure 8, bottom-right image) most of the attention previously focused on the group of people moving fast in the middle of the scene is inhibited.

Finally, a third behavior of top-down information can be seen on Figure 9. If some of the noise in the bottom-up attention map (Figure 9, top-right image) is inhibited on the top-down attention map (Figure 9, bottom-right image), there is also an area which is better highlighted (arrow). This is due to the fact that by inhibiting some of the motion perceived, the bottom-up process which consists in counting the number of blobs (section 3.2) can also find less of blobs, which mean a higher attention.



Fig. 9. Top-left: original frame, Top-right: bottom-up attention map (and also top-down attention map during the first viewing seconds), Bottom-left: the medium speed/high to-down model, Bottom-right: top-down attention map after inhibition by the models.

This section shows that top-down attention can have several kind of influences on bottom-up attention: from an inhibition of some areas, to the highlighting of other areas and passing by a negligible influence.

6 Conclusion

Behavior analysis of dense crowds has can benefit a lot from the use of models of human attention. 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 saliency index which models part of the human attention was demonstrated on dense crowd videos. After a first step of motion decomposition according to its speed and density, a bottom-up attention and top-down attention approaches were used. The rarity of the different components of this velocity and density decomposition is used to highlight the bottom-up abnormal or surprising behaviors. Repetitive motion in specific regions provides top-down information which is able to modify the final attention map. The results are surprisingly close to a human observer perception despite the fact that only speed-related features 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, additional evaluation of the algorithm on synthetic and real-life data is also required.

Acknowledgements

The research program “NUMEDIART” (www.numediart.org) is funded by the Walloon region, Belgium. Video databases from the “SERKET” project funded by the EU were used in algorithm tests.

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] V. Navalpakkam, and L. Itti, "Modeling the influence of task on attention", Vision Research, Vol. 45, No. 2, pp. 205-231, 2005.
- [9] 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.
- [10] M. Mancas. "Relative influence of bottom-up and top-down attention," Attention in Cognitive Systems, LNCS, Volume 5395/2009:pp. 212–226, 2009.
- [11] C. G. Healey, K.S. Booth, and J.T. Enns. "High-Speed Visual Estimation Using Preattentive Processing", ACM Transactions on Human Computer Interaction 3(2), pages 107-135, 1996.
- [12] RC Atkinson. Shiffrin. RM. "Human memory: A proposed system and its control processes", The psychology of learning and motivation: Advances in research and theory, 2:89–195, 1968.
- [13] K. Grill-Spectora, R. Henson and A. Martin, "Repetition and the brain: neural models of stimulus-specific effects", Trends in Cognitive Sciences, Volume 10, Issue 1, Pages 14-23, 2006.
- [14] M. Mancas, D. Glowinski, G. Volpe, P. Coletta, A. Camurri, "Gesture Saliency: a Context-aware Analysis", Springer Verlag, LNCS, to appear.