TASK-DEPENDENT SALIENCY ESTIMATION FROM TRAJECTORIES OF AGENTS IN VIDEO SEQUENCES

D. Campo[1], M. Baydoun[1] L. Marcenaro[1], C. S. Regazzoni[1,2]

Department of Electrical, Electronic, Telecommunications Engineering and Naval Architecture - DITEN, University of Genova, Italy[1]. Carlos III University of Madrid, Spain[2]

ABSTRACT

This paper proposes a method for detecting zones of visual attention based on the motion of agents in a video analytics context. By considering a Hough transform approach, linear flow motions are grouped based on attractive salient zones where they converge. Each group of linear flows is generalized through the whole environment by using a non-parametric stochastic approach that can be used to generate a map that illustrates the effects that each zone exerts on the dynamics of agents. A dataset of walking pedestrians and trajectories generated by a robot that executes a single task in a close environment are used to validate the proposed method.

Index Terms— Attention points detection, Gaussian process regression, trajectory analysis, vector field representation, Hough transform

1. INTRODUCTION

Visual attention refers to the process of orienting the gaze towards objects of interest inside an environment [1–3]. In everyday life, we are exposed to multiple environments composed of diverse types of objects. However, we only focus our attention on some of them. As suggested by [4], the ability to recognize quickly and efficiently salient objects in a scene has an evolutionary significance since it facilitates the localization of preys, mates or potential predators. Additionally, recent approaches [5] suggest that gaze plays a key role in the identification and importance that humans assign to objects.

Studies about gaze fixation under a human cognitive framework emerged from the studies of Yarbus [6], where people eye movements are analyzed while they are asked for characteristics of images. Former experiments related to human visual attention have been proposed by [2, 3]. In their works, they measure the time needed for identifying important parts of images under several conditions. Such experiments have motivated the study of saliency maps, which are representations of important regions in a certain image.

A formal definition of saliency maps comes with the work of Koch and Ullman [7]. They define saliency maps as a way to measure global conspicuity through the evaluation of multiple features. In their work, characteristics related to image properties such as color, direction of motion between frames, depth and orientation are taken into consideration. From this perspective, in a traditional sense, the saliency at a particular location is based on how image properties vary compared to its surroundings. In other words, the classical saliency can be seen as a measure of "novelty" or "surprise", where visual attractors are zones that pop-out from their vicinities [8].

As pointed out by [4, 9], visual attention can be explained from two viewpoints: i) bottom-up, image-based saliency cues and ii) top-down, task-dependent cues. In the first type, visual attention corresponds to the traditional saliency proposed by [7], where relevant zones are parts of the scene that are easily/quickly identifiable [4]. Visual attention of the second class, i.e., top-down way, depends on the accomplishment of a goal, which implies a determined task and a context, e.g., if we are asked to find a specific color or shape on an image. Such type of attention based on cognitive goals does not necessarily coincide with the traditional saliency output [10].

Although top-down task-dependent visual attention requires a voluntary "effort" from subjects for identifying specific spots in a scene [4], it is found to be a reasonable approach to understand human cognition since people are constantly immersed in contexts that make them focus on specific parts of their surroundings. An ideal way to model visual attention would be by combining both, bottom-top and top-bottom cues. This would facilitate the identification of salient areas that carry information of the contextual tasks that subjects execute. However, although top-down attention is a broad and powerful approach to understand cognition, most of research is focused on bottom-top strategies [11–15].

Motivated by the lack of research in top-bottom strategies for understanding visual attention, this paper introduces a novel proposal that aims at the identification of task-dependent salient zones based on the dynamics of subjects moving through a monitored environment.

The remainder of the paper is organized as follows: Section 2 presents the proposed methodology to find attention visual cues in videos, section 3 shows the results at evaluating our methodology in real scenarios and conclusions are presented in section 4.

Carlo Regazzoni has contributed to produce this work partially under the program "UC3M-Santander Chairs of Excellence."

2. METHODOLOGY

As proposed in previous works [16,17], it is considered a random walk dynamics as a baseline model to explain motions of agents inside an environment. Such model can be described as follows:

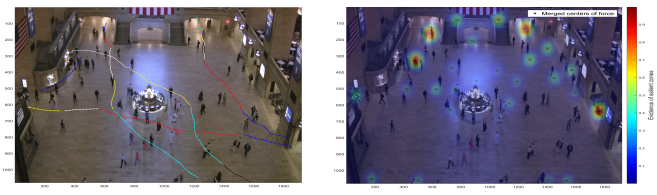
$$X_k = FX_{k-1} + w_{k-1}, \tag{1}$$

where X_k represents the state of each moving agent defined as its position and velocity at a time instant k, i.e., $X_k = [x_k \quad y_k \quad \dot{x}_k \quad \dot{y}_k]^T$. $F = [A \quad 0_{2,4}]^T$, $A = [I_2 \quad 0_{2,2}]$. w_k represents a zero mean Gaussian noise with covariance Q_k, i.e., $w_k = \mathcal{N}(0, Q_k)$. Q_k is here defined based on possible little motions that an agent with no particular motivation/task can perform. In other words, the model presented in (1) describes entities that stay in a single position without any particular tendency.

Agents that follow model (1) are here assumed not to be influenced by any particular part of the environment. Accordingly, deviations from a random walk model can be used to identify and characterize influences that environmental zones can exert over moving agents. By accepting assumptions about noise and dynamics of (1), it is possible to use a Kalman filter (KF) for tracking agents and take its innovations $\tilde{Y}_k^0$ to estimate properties of the scene.

$$\tilde{Y}_k^0 = Z_k - H_k X_{k|k-1}^0, \tag{2}$$

where $H = \begin{bmatrix} I_2 & 0_{2,2} \end{bmatrix}$, Z_k represents the measurements of trajectories and $X_{k|k-1}^0$ is the KF prediction based on a random walk model. Innovations $\tilde{Y}_k^0$ distant from zero are grouped into linear segments by using the method proposed in [17]. Fig.1 (a) shows how innovations produced by trajectories are decomposed into several linear segments.



(a) Segmentation process (b) Algorithm 1 output

Fig. 1. Segmentation and algorithm 1 results.

The last point of each segment is considered as a potential attractive point, i.e, a salient point candidate where agents tend to go. Attractive points are merged dynamically by an accumulative methodology shown algorithm 1.

In algorithm 1, r_t is a distance threshold with which potential attractive points are merged. The value of r_t is adjusted based on the distance between scene and surveillance camera. Such parameter affects the number of salient zones produced at finalizing the process. A high/low number of salient zones will be generated if r_t is decreased/increased respectively.

Algorithm 1 Identification of salient zones

Input:

$[r_t]$ Tolerance distance to merge centers of force
$i = 0$ Counter of ending segments
$l = 0$ Counter of merged salient zones

Output:

$[\overline{C}_l]$ Locations of estimated salient zones
$[n_l]$ Number of merged centers of force
procedure SALIENT ZONES DETECTION
loop:
    Wait for the ending of a segment
    $i = i + 1$
    $[c_x, c_y] \leftarrow$ Last point of the segment.
    if $i == 1$ **then**
        $l = l + 1$
        $C_l \leftarrow [c_x, c_y]$
        $n_l = 1$
        $\overline{C}_l \leftarrow [c_x, c_y]$
    else
        $r_l \leftarrow$ Vector of distances between $[c_x, c_y]$ and salient zones' locations $\overline{C}_l$.
        if $sum(r_t > r_l) > 0$ **then**
            $m \leftarrow$ Index of the closest zone to $[c_x, c_y]$.
            Append $[c_x, c_y]$ *to* C_m.
            $\overline{C}_m = mean(C_m)$
            $n_m = n_m + 1$
    goto *loop*.

By grouping dynamically potential attractive points, algorithm 1 enables to estimate a total number of P task-dependent salient zones, each of them indexed with $l \in 1, 2, ..., P$. n_l values can be seen as the result of a voting process that measures the evidence of each estimated salient zone centered in $\overline{C}_l$. Accordingly, high values of n_l represents a high validity of the salient zone indexed with l.

Fig. 1 (b) shows the output of algorithm 1. n_l values are normalized and used as a background color indicator that shows the evidence of each identified salient zone. Each center point $\overline{C}_l$ represented with a black cross can be seen as a center of attractive force [17] generated by each salient zone.

Assuming that pedestrians move in a quasilinear way towards salient zones they are looking at, it is proposed to cluster segments that point towards a common zone by using a Hough transform approach that incrementally calculates if a segment belongs to any previously identified salient zone. Algorithm 2 shows how this process is done.

A set of segments S_l associated to a salient zone centered in $\overline{C}_l$ is obtained by defining a tolerance distance d_t from its center of attractiveness. Parameters d_t and r_t are strongly correlated because both define how data is spatially grouped inside the scene. Accordingly, it is proposed to make $d_t \approx r_t$ to keep spatial symmetry between both algorithms.

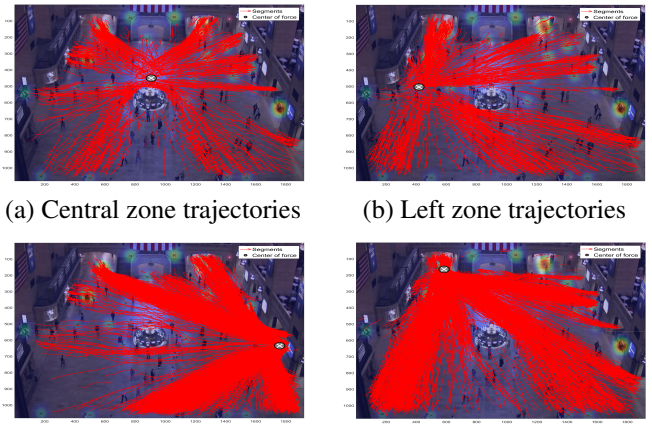
Algorithm 2 Assignation of segments to salient zones

Input:

$[\overline{C}_l]$ Locations of estimated salient zones
$[d_t]$ Tolerance distance for assigning segments
$[P]$ Total number of identified salient zones
$i = 0$ Counter of ending segments

Output:

$[S_l]$ Vectors of segments assigned to salient zones $\overline{C}_l$
procedure SEGMENTS ASSIGNATION
    Initialize S_l *as empty, where* l *goes from* 1 *to* P
loop:
    Wait for the ending of a segment
    $s_i \leftarrow$ Last completed segment.
    $i = i + 1$
    $L_i \leftarrow$ Linear model that fits the innovations of s_i.
    $d_l \leftarrow$ Vector of minimum distances between L_i and each center of force$\overline{C}_l$.
    if $sum(d_t > d_l) > 0$ **then**
        $M \leftarrow$ Vector of zones where d_l is less than d_t.
        Append s_i *to* S_M*vectors*
    goto *loop*.



(a) Central zone trajectories (b) Left zone trajectories

(c) Right zone GP (d) Up-left salient trajectories

Fig. 2. Algorithm 2 output: Clustered trajectories.

By considering that each segment is made of innovations $\tilde{Y}_k^0$ with their respective measurement Z_k, it is possible to define $S_l = [Z_{j(l)}, \tilde{Y}_{j(l)}^0]$, where $j(l)$ is the total number of innovations associated to the salient zone centered in $\overline{C}_l$. Fig. 2 shows the output of algorithm 2 for 4 identified salient zones.

A Gaussian process (GP) that takes $Z_{j(l)}$ as inputs and $\tilde{Y}_{j(l)}^0$ as outputs is considered for generalizing the effect that each salient zone exerts along the scene. Results from GP regression are included as a control input term $G^l(Z_{k-1})$ inside the random walk model as shown below:

$$X_k = X_{k-1} + UG^l(Z_{k-1}) + w_{k-1}, \tag{3}$$

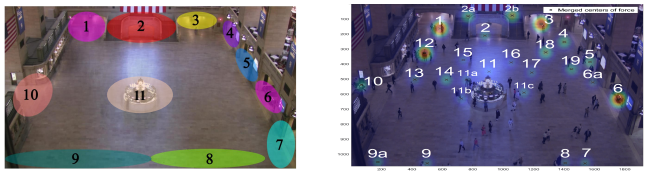
where $U = [I_2 \Delta k \quad I_2]^T$. Δk is the sampling time with which observations are measured. $G^l(Z_{k-1})$ is a nonlinear stochastic function obtained through a GP regression related to the salient zone centered in $\overline{C}_l$. l labels the detected salient zones. Assuming that a total of P salient zones are identified, it is possible to evaluate at each time instant the probability that any of them acts on a moving agent. For future work, the model in equation (3) can be used to improve the agents' tracking by considering their interactions with salient zones.

3. RESULTS

For validating the proposed method, 2 cases of study are considered: i) A pedestrian dataset taken from a fixed surveillance camera placed on the top of a real indoor environment. ii) Trajectories from a robot that accomplishes specific task.

3.1. Pedestrian Case

A dataset proposed by [18] is used to validate our method. The database is composed by 12,684 manually labeled walking routes taken from a surveillance camera. As suggested by [18], the environment can be decomposed in ten global source/destination zones depicted in Fig.3 (a) from 1 to 10. Additionally, an 11$^{\text{th}}$ central zone is considered as other point that exerts a momentary attraction to some pedestrians.



(a) Ground truth from [18] (b) Detected salient zones

Fig. 3. Task-dependent attractive visual zones in the scene

Fig. 3 (b) is a labeled version of salient zones identified by applying algorithm 1. It is possible to see that a total of 26 saliency zones are identified through the proposed methodology. By comparing found zones with Fig. 3 (a), it can be seen that some of them are related to single attractive zones defined in Fig. 3 (a), it is the case of zones 2, 6, 9 and 11, which have more than a single assignation in Fig. 3 (b).

By comparing Fig. 3 (a) with Fig 3 (b), it is possible to see that more detailed information can be obtained with the proposed method, e.g., for zone number 2, the proposed methodology finds 3 destination points (2, 2a and 2b) instead of a single one. Additionally, our method identifies task-dependent places, it is the case of zone 12, which corresponds to a queue point. Others zones, i.e., from 13 to 19, correspond to places where pedestrians tend to go due to common tasks that they execute, e.g., going from one door to another one.

By computing a GP regression of horizontal/vertical innovation components produced by a particular salient zone over

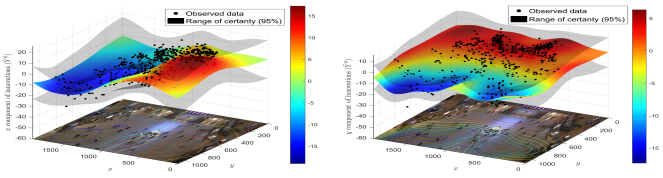
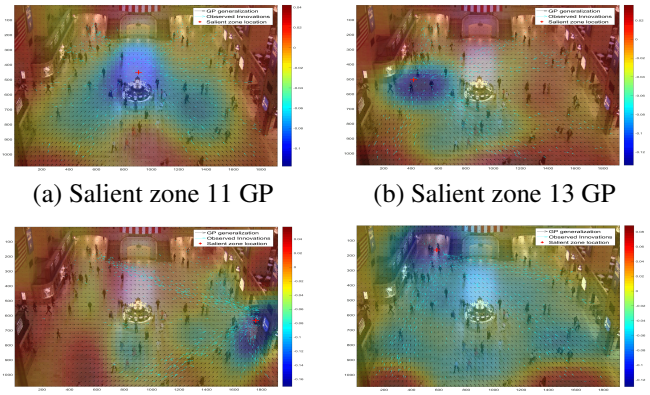


Fig. 4. GP regression in both components of innovations produced by salient zone 11.

all the scene, it is possible to obtain a pair of stochastic maps generated by the zone in question such as shown in Fig. 4.

Fig. 4 shows the couple-map of innovations generated by the salient zone 11. By considering the mean value of each map (shown as a color-map), it is possible to build a vectorial map of the most probable innovations that a particular salient zone will produce through all the scene. Fig. 5 show the vectorial map of innovations obtained for 4 different salient zones out of the 26 that were identified.



(a) Salient zone 11 GP (b) Salient zone 13 GP

(c) Salient zone 6 GP (d) Salient zone 1 GP

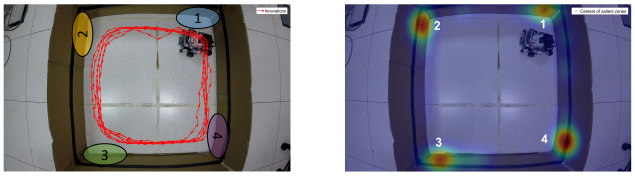
Fig. 5. Divergence maps using pedestrians GP regressions.

Background colors in graphics of Fig. 5 are related to the divergence of innovation vector fields obtained by the GP regression. Accordingly, low values (blue areas) coincide with locations of attractive salient zones and high values (red areas) are related to zones from which people come from when they are going towards the saliency zone in question. From that viewpoint, graphics of Fig. 5 can be seen as radiation maps that show each saliency zone's effects over all the scene.

3.2. Robot Case

A robot with the task of turning in one sense inside a closed scene as shown in Fig. 7 (a) is monitored with a fixed camera.

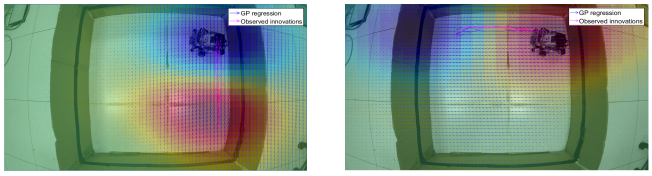
Fig. 7 (a) shows the expected ground truth of the proposed experiment. As can be seen, a turn can be seen as a combination of effects exerted by 4 task-depended salient zones. By applying algorithm 2 over the robot trajectories, segments that behold to each salient zone are correctly identified.



(a) Ground truth (b) Detected salient zones

Fig. 6. Robot case layout.

As done in the pedestrian case, a vector map that explains the effect of each salient zone is approximated by a GP regression that uses as inputs identified innovation segments pointing towards the zone question. Fig. 7 shows results of zones 1 and 2. It can be seen how divergence maps in the background can explain transitions between zones. Accordingly, Fig. 7 (a) shows that trajectories attracted by zone 1 come from the zone 4. Similarly, Fig. 7 (b) shows that trajectories influenced by zone 2 come from the zone 1.



(a) Salient zone 1 GP (b) Salient zone 2 GP

Fig. 7. Divergence maps using robot GP regressions.

4. CONCLUSIONS

It is proposed and validated a methodology for extracting task-dependent visual attractive cues in two different real scenarios. Results suggest that the proposed method can be used in different video data in which activities of moving agents can be simplified in simple tasks, such as going from a particular point to another one.

Maps obtained in this work are proposed to be used as complementary information for traditional saliency methods. Since video data that captures dynamics of agents inside a scene is available, it is possible to analyze their movements to identify important zones in a scene. Results of this work can be seen as information that is not necessarily correlated with the results that can be reached with traditional saliency methods.

By applying a Gaussian process regression based on sparse observations, it is possible to estimate the effect of salient zones through a whole scene. Such information can be used in future works to improve the capability of prediction by introducing relations of causality between salient zones and moving agents.

5. REFERENCES

[1] W. James, *The Principles of Psychology*, Number v. 1 in American science series. H. Holt, 1890.

[2] A.M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[3] M.I. Posner, "Orienting of attention.," *The Quarterly journal of experimental psychology*, vol. 32, no. 1, pp. 3–25, 1980.

[4] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[5] Niki Martinel, Christian Micheloni, and Gian Luca Foresti, "Kernelized Saliency-Based Person Re-Identification Through Multiple Metric Learning," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5645–5658, dec 2015.

[6] A.L. Yarbus, *Eye movements and vision*, Plenum Press, New York, 1967.

[7] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[8] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, no. 10, pp. 1295–1306, 2009.

[9] S.-H. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, "Video saliency detection via dynamic consistent spatio-temporal attention modelling," 2013, pp. 1063–1069.

[10] D.H. Ballard and M.M. Hayhoe, "Modelling the role of task in the control of gaze," *Visual Cognition*, vol. 17, no. 6-7, pp. 1185–1204, 2009.

[11] M. Hua and Y. Li, "A novel visual saliency detection method using motion segmentation," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 11, no. 5, pp. 417–424, 2016.

[12] J. Li, F. Meng, and J. Mao, "Saliency detection on videos with scene change," 2015, pp. 506–510.

[13] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 171–177, 2010.

[14] K. Fragkiadaki, G. Zhang, and J. Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," 2012, pp. 1846–1853.

[15] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision Research*, vol. 47, no. 19, pp. 2483–2498, 2007.

[16] D. Campo, V. Bastani, L. Marcenaro, and C. Regazzoni, "Incremental learning of environment interactive structures from trajectories of individuals," in *2016 19th International Conference on Information Fusion (FUSION)*, July 2016, pp. 589–596.

[17] Damian Campo, Alejandro Betancourt, Lucio Marcenaro, and Carlo Regazzoni, "Static force field representation of environments based on agents' nonlinear motions," *EURASIP Journal on Advances in Signal Processing*, vol. 2017, no. 1, pp. 13, 2017.

[18] Shuai Yi, Hongsheng Li, and Xiaogang Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015.