Statistical Analysis of Visual Attentional Patterns for Video Surveillance

Giorgio Roffo[†], Marco Cristani[†⊕], Frank Pollick[‡], Cristina Segalin[†], and Vittorio Murino[⊕†]

[†] Department of Computer Science, University of Verona (IT)
[⊕] Pattern Analysis and Computer Vision Dept., Istituto Italiano di Tecnologia (IT)
[‡] School of Psychology, University of Glasgow (UK)

Abstract. We show that *the way* people observe video sequences, other than *what* they observe, is important for the understanding and the prediction of human activities. In this study, we consider 36 surveillance videos, organized in four categories (*confront, nothing, fight, play*): the videos are observed by 19 people, ten of them are experienced operators and the other nine are novices, and the gaze trajectories of both populations are recorded by an eye tracking device. Due to the proved superior ability of experienced operators in predicting violence in surveillance footage, our aim is to distinguish the two classes of people, highlighting in which respect expert operators differ from novices. Extracting spatio-temporal features from the eye tracking data, and training standard machine learning classifiers, we are able to discriminate the two groups of subjects with an average accuracy of 80.26%. The idea is that expert operators are more focused on few regions of the scene, sampling them with high frequency and low predictability. This can be thought as a first step toward the advanced automated analysis of video surveillance footage, where machines imitate as best as possible the attentive mechanisms of humans.

Keywords: surveillance, gaze control, eye movement analysis, activity recognition, eye tracking

1 Introduction

The study of eye movements is an innovative way of assessing the skill in monitoring of Closed Circuit Television (CCTV) recording, in which a comparison of the eye movement strategies between experienced operators and novice observers may show important differences that could be used in training an automatic monitoring system. Generally, when we are looking at a video, we consciously or unconsciously focus only on a fraction of the total information that we could potentially process, in other words we perform a perceptual selection process called *attention*. Visually, this is most commonly done by moving our eyes from one place of the visual field to another; this process is often referred to as a change in *overt attention* – our gaze follows our attention shift. The process of selecting

visual information is crucial for the subsequent activity understanding, where internal mental representations are built for categorizing the observed events and starting to reason on them, for example to predict future actions.

In this paper, we focus on extracting the spatio-temporal eye patterns which regulate the attentive processes of experienced operators, looking if they substantially differ from those of novice people. Due to the higher ability of experienced operators in predicting violence in surveillance footage [12], we argue that understanding the way visual information is processed can be important for automated video surveillance.

Our approach aims at individuating where the focus of attention is located on the scene and the dynamics of this process. Considering gaze trajectories and modeling them in diverse fashions (e.g., encoding local curvatures, feeding them into heterogeneous classifiers as [6], etc.) did not reveal in our experiments significant differences between experts and novices. Therefore, we follow another strategy, which focuses on two different logical layers, spatial and temporal. Spatial analysis is performed by analyzing the zones of the screen considered most of the time: partitioning the image into cells and counting how many times they have been watched, indicates strongly different patterns among the two classes of observers. For the temporal characterization, we analyze the unpredictability of the movement patterns by adopting entropic measures, capturing in practice the irregularity of the eye trajectories. Spatial and temporal analyses are carried out with standard classifiers (SVM and kNN, respectively), and the fusion of the classification results allows one to consistently separate experts from novices, with an accuracy of 80.26%. In particular, we find that experts are characterized by a spatially more focused analysis (*they know where to look*) with a high level of unpredictability (basically, they switch continuously among different spatial cells), while novices tend to show more regularity in the analysis, considering a larger area of analysis, with a lower speed in accessing the data.

The rest of the paper is organized as follows. In Sec. 2, a review of the related literature is presented, and Sec. 3 details the proposed approach. Experiments are reported in Sec. 4, and, finally, Sec. 5 draws some conclusions and future perspectives.

2 Related work

The selection of good CCTV operators is essential for effective CCTV system functioning. The study of gaze control mechanism is an intriguing way for evaluating the skills of entry level CCTV operators. Indeed, how gaze control operates over complex real-world scenes has recently become of central concern in several core cognitive science disciplines including cognitive psychology, visual neuroscience, and machine vision. For example, an application of psychological principles to Aviation Safety and Welfare (ASW) is suggested in [8], which analyzes the eye movements of expert and novice pilots while performing landings in a flight simulator. They found that expert pilots had significantly shorter dwells, more total fixations and they observe a specific place of interest in the visual

scene. Experts were also found to have better defined eye-scanning patterns. In [11], authors conducted a comparison of the eye movement strategies between expert surgeons and novices, while performing a task on a computer-based laparoscopic surgery simulator: the results from eye gaze analysis showed that experts tended to maintain eye gaze on the target, whereas novices were more varied in their behaviours. In general, gaze control differs during complex and well-learned activities such as reading [14], tea and sandwich making [9], and driving [10].

Going back to surveillance, an ongoing research programme is investigating the ability of humans to detect whether or not an individual, captured on CCTV, is carrying weapons [5]. In [2], trained CCTV operators and lay people viewed footage material and were asked to indicate whether or not they thought the surveillance target was carrying a firearm. Our work is in line with this type of research.

3 Our approach

Our approach partitions the screen in a set of 5×5 non-overlapped squared cells, of size 288×180 pixels each. From this support, we calculate two sets of features: the former models explicitly *where* the attention of the subject has been driven during the monitoring activity, and we call it *spatial feature set*. The latter indicates *how* the attentional analysis has been performed by the subjects, and we call it *temporal feature set*.

The spatial feature set is composed by one feature, which is the **Cell Counting (*Count*)**: a counting matrix, where the i^{th} cell records exactly how many times a participant has been watching the i^{th} cell of the grid. In practice, each videosequence can be summarized by a 25-dim count vector.

In the temporal feature set, the features have been designed upon three temporal basic cues that we will present below. The idea is that eye movement information is recorded, storing for each i-th cell a number $f(i)$ of basic cue values, where $f(i)$ indicates the number of times the i-th cell has been intercepted by an eye trajectory.

Three are the temporal basic cues:

- **Fixation Duration (*FIXd*)**: a fixation is the state of the eyes during which gaze is held upon a specific region. Humans typically alternate saccadic eye movements and fixations. The term "fixation" can also be referred to as the time between two saccades, during which the eyes are relatively stationary [7, 16]. In our experiments, for each video analyzed by a subject, the time spent for each fixation in a particular cell has been recorded, expressed in ms. Therefore, for each cell we have a sequence of fixation duration values.
- **Saccades Velocity (*SACv*)**: the eyes do not remain still when viewing a visual scene; they have to move constantly to build up a mental "map" from interesting parts of the scene. The main reason for this is that only a small central region of the retina, the fovea, is able to perceive with high acuity. The simultaneous movement of both eyes is called a saccade. The duration

of a saccade depends on the angular distance the eyes travel during this movement, the so-called saccade amplitude. A saccade is individuated as a movement exceeding the threshold of $\tau = 30°/sec$ starting after the fixation, lasting at least 20 ms [15, 1]. For each cell we record all the saccades′ related speed values calculated over it, measured in *degrees/seconds*.

– **Smooth Pursuit Velocity (*PURv*)**: smooth pursuit is the eye movement that results from visually tracking a moving object. Generally, this kind of eye movement has a speed lower than $30°/sec$ [13, 16]. The PURv is measured in *degrees/seconds* and the values are stored as for the previous cues.

In practice, as description of the whole monitoring analysis performed on a video sequence by a subject, we obtain three different cue volumes, each related to the $FIXd$, $SACv$ and $PURv$ feature. In the $i-$th entry of each volume we have all the $f(i)$ feature values collected in the $i-$th cell (i.e., depending on how many times that cell has been visited). At this point, to obtain a unique cue value for each $i-$th entry, we applied the mean operator. As a result, we obtained the 5×5 maps μ_x, where x stands for $FIXd$, $SACv$ and $PURv$.

At the end, in order to distill a single measure from each map, we calculate its *entropy*: this way, we obtained three entropic values for each analyzed videosequence, dubbed E_{FIXd}, E_{SACv} and E_{PURv}. The underlying rationale of choosing entropic measures consists in the fact that the entropy gives a measure for assessing how unpredictable is the behavior of the subject: high entropy means that in the whole sequence the subject behaved in a very dynamic fashion, for example steadily focusing on some scene details, then suddenly moving the focus of attention toward distant screen locations. Viceversa, low entropy indicates that the subject kept repeated attentional patterns, patrolling in a mechanical fashion the screen.

Spatial and temporal features become the signature of the attentive behaviour of a single subject: given a pool of subjects belonging to the same class, our approach learns a classifier by employing linear Support Vector Machines (SVM) on the 25-dimensional spatial features, while the 3-dimensional temporal features are processed by kNN classifiers. The choice of the classification machinery supported us with satisfying results, as witnessed in the next section.

4 Experiments

In the experiments, we apply our approach to a recent video dataset provided by the University of Glasgow, whose content is detailed in the following.

4.1 The dataset

The dataset has been taken from tens of urban surveillance cameras, highlighting "hot zones", that is, crossroads near pubs and discotheque areas. In particular, thirty-six 16-second CCTV clips were used. These videos have been grouped in four categories (see Table 1), each composed by 9 videos: in the "Fight" category,

behaviours leading up to a violent incident are shown; in the "Confront" category, a sequence of behaviours similar to the fight clip are shown, although no violent/harmful incident occurred; the "Play" category shows people interacting in a playful manner; finally, the "Nothing" category includes a variety of scenes where no violent/harmful behaviour occurs and they were taken from similar locations and with similar camera views. Please note that in the experiments, videos of the Fight category have been truncated, so that fights are not visible: this design was necessary to highlight solely the attentional behavior needed to understand the situation and predict the outcome, and not to analyze the outcome itself. The eye tracking experiment was attended by 19 participants, 10 CCTV operators (3 female, 7 male) aged 21-53 years ($\mu_{age} = 36.3$, $\sigma_{age} = 10.1$); and 9 novices (2 female, 7 male) aged 28-43 years ($\mu_{age} = 33.8$, $\sigma_{age} = 6.0$). All participants were native English speakers, naïve to the goals of the experiment and had not participated in eye tracking experiments in the past. All the participants had normal binocular (Titmus Test) and colour vision (CUCV Test) and corrected binocular visual vision acuity of 6/9 or better. Three of the participants wore eye glasses during the experiment, and two wore contact lenses. The device was an ASL Eye-Trac6 remote eye tracking device, located directly below the display screen and 0.65 meters from the participant's eye. A chin rest was used to minimise head movement and to maintain viewing distance. The video were displayed on a 19 inch LCD monitor with a set resolution of 1440×900 pixels which described a $37^\circ \times 23^\circ$ field of view.

Fight clip	Behaviours leading up to a violent incident.
Confront clip	Confronts which did not lead to a fight.
Play clip	People interacting and some playful encounter happens.
Nothing clip	Scenes where no violent/harmful behaviour occurs, taken from similar locations and with similar camera pans.

Table 1. Categories of CCTV clips. A violent incident was defined as an aggressive physical contact with intent to harm, such as a slap, shove, punch, or kick.

As preliminary analysis of the dataset, basic statistical analysis on standard features has been carried out. In particular, we consider the *mean fixation time* as the percentage of time a subject spends fixating when viewing the clip, the *mean fixation duration* as average duration of all the fixations on a given video and the *mean saccade rate* as the average number of saccades made per second. A main difference among clip categories was observed for the eye movement measures of gazing time and fixation duration. It indicates that there were significant differences in participants′ gazing time and fixation duration when viewing different types of clips. In particular:

– Participants exhibited significantly longer gazing time for clips in the matched confront clip category ($\mu = 80.08$, $\sigma = 3.66$), when compared to fight clips

($\mu = 78.31$, $\sigma = 3.99$, $p = 0.008$), to play clips ($\mu = 74.54$, $\sigma = 4.78$, $p < 0.001$) and to nothing clips ($\mu = 76.37$, $\sigma = 4.34$, $p < 0.001$).
- Although not statistically significant, a trend was found that CCTV operators spent lower proportion of time making fixations ($\mu = 76.16$, $\sigma = 4.19$) when compared to novice participants ($\mu = 78.5$, $\sigma = 4.8$). This may suggest that CCTV operators spent more time engaged in saccades and/or smooth pursuit tracking during the clip than novices.
- The mean fixation duration data revealed that CCTV operators exhibited a shorter mean fixation duration ($\mu = 0.34$, $\sigma = 0.02$) in comparison to novice participants ($\mu = 0.36$, $\sigma = 0.04$), even if this difference was not statistically significant.
- A third test was conducted to investigate if there were any significant differences in the mean rate of saccades due to participant experience. This analysis found no main effect of experience.

These results highlight differences between the two groups but do not explain what was observed by the subjects and in what way this happened.

4.2 Results

The goal of the classification was to divide novice people from expert operators and this was performed in the following way. First of all, we separate the analysis carried out on the spatial and the temporal features, to assess the contribute of each group of cues. In all the cases, Leave-One-Out cross validation was performed, considering a particular subject as test element, keeping the others as training samples, and exploring all the possible training/test partitions, averaging the classification values at the end. Since each subject watched 9 videos, we build 9 classifiers, i.e., one for each video. Given a test subject, we evaluate its "novice" or "expert" label by majority vote, considering the results of the 9 classifiers. For the *Count* spatial feature, we employ linear SVM as classifier, while for the entropic temporal features E_{FIXd}, E_{SACv} and E_{PURv} we adopt the kNN algorithm. The choice of these classifiers gave us the best performances, and their parameters have been chosen by cross-validation.

In the spatial analysis, some *Count* counting matrices have been reported in Fig. 1. Qualitatively, one can see that expert operators are more focused on a central smaller area (which collected the highest number of votes) while novices are more spread over the entire image plane. It is worth noting that these areas were populated by human subjects[1]. The quantitative results are reported in Tab. 2.

In the case of the entropic temporal features, for each subject we considered 9 kNN classifiers, one for each video. The results were quite higher than the spatial counterpart (Tab. 2). For evaluating the effect of including both the spatial and temporal features in the classification process, the majority vote was applied to the all the 18 classifiers, 9 for the spatial features and 9 for the temporal features. We do the same strategy for all the 19 subjects, averaging at the end

[1] The footage cannot be shown for ethical and privacy issues.

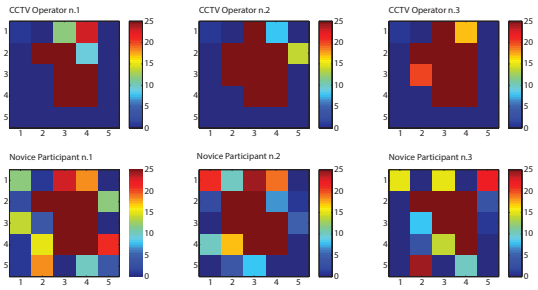

Fig. 1. Spatial analysis of the *Count* matrices. The figure shows that the CCTV operators focus on smaller areas than the novices.

Activity/Features	Temporal	Spatial	Joint
Fight	78.9%	68.4%	84.2%
Play	63.1%	73.7%	63.2%
Nothing	84.2%	78.9%	84.2%
Confront	73.7%	68.4%	89.5%
Average	75.0%	72.3%	80.3%

Table 2. Classification rates while considering Temporal and Spatial cues separately and jointly (third column).

the accuracy scores obtained for each person. The results are shown in Table 2. We noted that:

- In general (apart from the *Play* class), temporal features were more effective in separating the two classes;
- In general (apart from the *Play* class), the fusion of spatial and temporal features was no worse than the single classifiers, showing a certain complementarity between the two different modeling schemes.

5 Conclusions

In this paper we presented an analysis which considers eye tracking data on video surveillance sequences. Our goal was to understand how expert CCTV operators analyze such videos, and if there is a difference with novice participants. Extracting spatio-temporal features, and training SVM and kNN classifiers, we have been able to discriminate the two groups of subjects with an average accuracy of 80.26%: the idea is that expert operators are more focused on few regions of the scene portraying the humans, sampling them with high frequency. This study follows the recent trend of applying a social signal processing perspective to surveillance [3, 4], where psychological analyses are exploited to inspire more effective monitoring strategies. In particular, this can be thought as a first step

toward the advanced automated analysis of video surveillance footage, where machines imitate as best as possible the attentive mechanisms of humans: in this case, the take-home message is that the dynamics with which people are observed is highly unpredictable but highly focused on them. Even if these results may appear intuitive, they have been obtained by a solid experimental analysis, for the first time.

References

1. A.T. Bahill, M.R. Clark, and L. Stark. The main sequence, a tool for studying human eye movements. *Math. Biosci*, (2), 1975.
2. A. Blechko, I. Darker, and A. Gale. Skills in detecting gun carrying from CCTV. In *International Carnahan Conference on Security Technology*, 2008.
3. M. Cristani, V. Murino, and A. Vinciarelli. Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In *CVPRW 2010*, pages 51–58, 2010.
4. M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, 100:86–97, January 2013.
5. G. Hales, C. Lewis, and D. Silverstone. *Gun Crime: The Market in and Use of Illegal Firearms.* Findings (Great Britain. Home Office. Research, Development and Statistics Directorate). Home Office, 2006.
6. J. M. Henderson, P. A. Weeks, and A. Hollingworth. Multi-feature object trajectory clustering for video analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1555–1564, 2008.
7. R. Ji, X. Sun, and H. Yao. What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, (3):1552–1559, 2012.
8. P. Kasarskis, J. Stehwien, J. Hickox, A. Aretz, and C. Wickens. Comparison of expert and novice scan behaviors during vfr flight. In *Proceedings of the 11th International Symposium on Aviation Psychology*, 2001.
9. M. F. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25-26):3559–3565, 2001.
10. M.F. Land and D.N. Lee. Where we look when we steer. *Nature*, 369:742 – 744, June 1994.
11. B. Law, M.S. Atkins, A.E. Kirkpatrick, and A.J. Lomax. Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. pages 41–48, 2004.
12. K. Petrini, P. McAleer, C. Neary, J. Gillard, and F.E. Pollick. Experience in judging intent to harm modulates parahippocampal activity: an fmri study with experienced cctv operators. In *European Conference on Visual Perception*, 2012.
13. J. Pratt. Visual fixation offsets affect both the initiation and the kinematic features of saccades. *Experimental Brain Research*, 118(1):135–8, 1998.
14. K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422, November 1998.
15. A. Torralba. Modeling global scene factors in attention. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 20(5):1407–1418, 2003.
16. A. Torralba, M. S. Castelhano, A. Oliva, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113, 2006.