Visual Attention and Memory Augmented Activity Recognition and Behavioral Prediction

Dr. Srikanth Prabhu[1] and Nidhinandana Salian[2]

Department of Computer Science Engineering, Manipal Institute of Technology, Karnataka, India
srikanth.prabhu@manipal.edu nidhisalian08@gmail.com

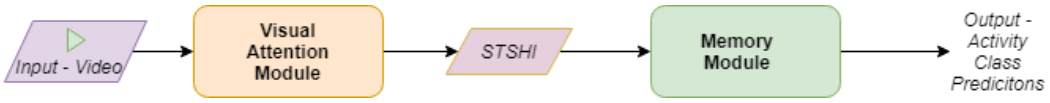
Abstract. Visual attention based on saliency and human behavior analysis are two areas of research that have garnered much interest in the last two decades and several recent developments have showed exceedingly promising results. In this paper, we review the evolution of systems for computational modeling of human visual attention and action recognition and hypothesize upon their correlation and combined applications. We attempt to systemically compare and contrast each category of models and investigate directions of research that have shown the most potential in tackling major challenges relevant to these tasks. We also present a spatiotemporal saliency detection network augmented with bi-directional Long Short Term Memory (LSTM) units for efficient activity localization and recognition that to the best of our knowledge, is the first of its kind. Finally, we conjecture upon a conceptual model of visual attention based networks for behavioral prediction in intelligent surveillance systems.

Keywords: visual attention, activity-recognition, behavioral-predictio

1 Introduction

Intelligent visual surveillance in dynamic real-world scenes is an upcoming arena in the field of computer vision and automation and endeavors to recognize human activity by learning from prerecorded image sequences, and more generally to understand and predict subject behavior. A key precursor to arresting suspicious activity, which is the ultimate goal of most intelligent surveillance systems, is the accurate detection of peculiar or irregular movement that is observably deviant to the norm. Attention enhances the rationality of the agent by implementing a bottleneck which allows only relevant information to pass to higher level cognitive components like object recognition, scene interpretation, decision making and memory for further processing.

This paper attempts to bring dynamic visual saliency to the fore as a potential technique for anomalous activity discovery and surveys existing methodology that could provide a reliable means to recognize the most significantly noticeable features in videos of illegal activity and thus provide the basis for a predictive model to recognize the beginning of an occurrence of a potential crime. In our study, we have presented a complete overview of a variety of approaches taken to tackle the problem of saliency detection, and we suggest a biologically plausible framework applicable to the task of activity recognition and behavioral prediction in natural envi-



ronments. Figure 1 is a diagrammatic representation of our model for saliency-based activity recognition.

Fig. 1. A Brief Overview of Our Model for Saliency-Based Activity Recognition

In order to understand dynamic saliency, it is important to explain the evolution of various lines of thought attempting to successfully model visual conspicuity in still images. The various approaches for saliency detection we have studied in our survey include the neuroscience-psychology based models that isolate prominent low-level features, task-specific models that serially scan a visual scene for target objects, reward-based reinforcement learning models, information maximization theory based on statistical rarity of local features and self-information, the computational-mathematical approach of regional covariance and the increasingly popular neural network based models. We have tried to present the advancement in this arena in chronological order as far as possible, presenting the pros and cons of each model and how subsequent developments attempt to meet the shortcomings of their predecessors.

The earliest models on behavioral prediction, on the other hand, were based on theoretic decision tree-like models of possible action sequences, assuming perfectly rational agents. Later models based on sample image sequences of human movement considered each agent capable of several internal states and introduced concepts such as mean gestures and gesture phases and used Dynamic Time Warping, Finite State Machines or Hidden Markov Models to explain transitions between them. More recent models introduced in the last two years have explored the idea of using various types of deep neural networks to learn predictive algorithms for human behavior with increasing accuracy. Most of these models use the given initial sequence of behavioral states or 'gestures' to output a predictive sequence of future states or gestures. Recent hypotheses have recently relabeled this task more appropriately as activity estimation, and the identification of the preparatory gestures as comparable to implicit activity recognition. Activity recognition is an important precursor to overall human behavioral analysis – because human behavior itself consists of several simultaneous activities. In our study, our primary focus will be on the activity recognition element of the existing paradigm, and how it could be integrated into visual attention prediction mechanisms for an end-to-end intelligent surveillance system for observable anomaly detection.

The remainder of this paper is structured as follows – Section II(A) illustrates the origins and progression of computational architecture modelling visual attention, the shortcomings and successes of subsequent models. Section II(B) presents a brief overview of the development and current dominant paradigms for behavioral prediction models and an introduction to the advancements in activity recognition models. In section III, we present a novel theory on a combined model of Saliency-Based Context-Aware Activity Recognition for implicit activity localization and identification using a pre-trained network for spatiotemporal saliency prediction, as a precur-

sor to the conceptual model of Saliency based Behavioral Prediction, discussed in Section IV. We also address the limitations faced by the model from Section III, which we are currently in the process of implementing, and the future work required for this approach to become a viable paradigm for intelligent surveillance systems.

2 Prior Work

2.1 Saliency Detection Models

Early theories attempting to explain visual saliency can be divided into two categories- top-down and bottom up, emulating the pre-attentive and attentive phases of attention in human cognition. While bottom up approaches were stimulus-driven and assumed that a spectrum of low level features in a scene were processed in parallel to produce cortical topographical maps, top-down features considered the approach that visual observation was goal-oriented and conducted by serially attending probable target locations. Koch and Ullman[1] were the first to recognize the similarity between the underlying neural circuitry and a hierarchy of elementary feature maps in the 80s, inspired by the pioneering work of Treisman et al. on feature integration theory[2]. The first working model of this proposed architecture was introduced by Itti et al[3]. It was later discovered that their winner-takes-all approach was neurally plausible and it was likely that there was a similar reward based mechanism used for memorization and learning within neurons in the brain[4-5] and many reinforcement learning models were proposed in abidance with this line of thought[6-7].More research by neurobiologists revealed a rapid decline in visual acuity while moving from the center of the retina(fovea) towards the periphery, and the need to rapidly realign the center of the eye with the region of focus by means of saccades[5][8]. This provided a plausible explanation for the center bias observed in human eye fixation data over static images, and later models that took this into account proved far more effective than earlier architectures[9-12]. This theory was later advanced by the belief in the existence of a deictic system in the brain for memory representation in natural tasks, in order to economize the internal representation of visual scenes[8][5][13-14]. Observations of individual eye movements in later experiments were found to suggest that information was incrementally acquired during a task and not at the beginning of a task[15].Similar studies also showed that the initial internal representations most likely included illusory conjunctions formed between non-attended locations[2][16] and was used to form a spatial ‘gist’ of the visual scene. This theory has recently been revisited by researchers attempting to model visual attention and memory[17].

Around the same period, the idea of a fast, automatic mapping of elementary features followed by serial scan of ‘interesting’ locations began to develop and many computational models around this time tried to replicate human eye fixations by using either objective, stimulus-driven or subjective, task-driven approaches[16][18-20].Thus the idea of selective visual attention was popularized and several decision-theoretic models based on sequential decision making, mostly emulating search

tasks, were introduced[21]. Information theoretic models published at around the same time introduced the influence of prior experience and self-knowledge as a factor in gaze allocation [22-23].

Advancements in neurobiological research around this time revealed the shared and specific neural substrates involved in the processing of perceptual identification (ventral stream) and spatial localization (dorsal stream) of target objects[24]. The conspicuity of spatially (pop out) and temporally (surprise motion) scarce factors was later thought to be an aspect of top down processing of visual scenes [25], through various research into inattentional blindness while performing a specific task[21]. Of late, a new notion that combines these two approaches has been suggested – introducing the idea of simultaneous covert and overt visual attention, stating that initially the brain creates a low level saliency mapping that takes into account the elementary features such as colour, orientation and intensity; and then conveniently uses that initial representation while pursuing a given task, oblivious to changes unless explicitly notified[26-27]. Several prior experiments into this arena have shown results consistent with this theory, many recent models even relying upon advancements in virtual reality technology to create realistic yet completely controllable models of natural environments[16].

Although most of the existing neural network models are trained on vast eye fixation data[28-31] for static image viewing in controlled environments, even using neural networks to learn hidden non-linear mappings between images and fixations[32-34], new research has shown that this is a potentially problematic paradigm that presents a stark contrast to human eye fixations in natural, uncontrolled environments[35]. Recent studies have shown that although humans tend to fixate on faces and text irrespective of the task while watching images and videos in a controlled environment[14][27][36], this is not the case in real-world interactions, presumably because the faces in context belong to other human beings who could possibly look back at the observer[35][37][38], and similar experiments in the field of human psychology corroborate this finding, revealing that this potential for social interaction could possibly cause one to experience an elevated level of self-consciousness[39][40]. This once again raises questions over the current state-of-the-art deep learning based models that use pre-trained object detection CNNs to fine tune saliency detection[41], on the assumption that humans predominantly tend to fixate on faces, text and objects[42-43]. Multi-scale[44-45] and multi-stream[46-50] convolutional networks and recurrent saliency detection models[34][51] that have been introduced relatively recently, appear to provide an acceptable alternative that could perform equally well in controlled and natural environments.

2.2 Behavioral Prediction Models

Foreseeing the conduct of human members in strategic settings is a vital issue in numerous areas. Early work in this arena involved mostly expert-constructed decision theoretic models that either expected that members are superbly rational, or endeavored to show every member's intellectual forms in view of bits of knowledge from subjective brain research and experimental economics[52].

Behavioral prediction and gesture analysis models, on the other hand, have seen rapid progress in the models that consider a human to be a complex mechanical device, capable of several internal (mental) states that may be indirectly estimated. The prevalent notion in this area of research is that behavior can be divided into a combination of sequential, interleaving and concurrent activities, which in turn can be represented as a combination of actions. There are two main monitoring approaches for automatic human behavior and activity estimation, viz., vision-based [53-54] and sensor-based[55-56] monitoring. In this study, we will focus primarily on the vision-based behavioral approximation models.

The very first models that tried to capture the temporal aspects of movement tried to combine images from multiple viewpoints at multiple instants of time and tried to create a binary feature vector in the form of some variation of a temporal template[57][58]. This traditional paradigm relied primarily on a collection of static image samples to calculate a mean gesture and then attempted to generalize this description to the rest of the data. The earliest models that set the precedent for prototype-based matching and tracking for gesture analysis that could clearly capture the smoothness and variability of human movement were based on Dynamic time Warping (DTW)[59] for temporal scale invariant action recognition and Finite State Machines (FSMs)[60], that considered gestural phases to be disjointed finite states, connected by transitions. Experimental results also converged with the theory that human actions are more suitably described as a sequence of control steps rather than as a combination of random positions and velocities. In order to decode states from observations, the idea of using Hidden Markov Models – a concept then being explored in text prediction and speech recognition models[61-62], eventually made its way to behavioral analysis as well. In later models, prototypical dynamic movements that form the basic elements of cortical processing were represented as a large set of linear controllers (eg. augmented Kalman filters), that could be sequenced together with a Markov network of probabilistic transitions to represent motion and behavior [63]. These were succeeded by models based on Conditional Random Fields (CRFs) [55][64] that could capture long range dependencies better than traditional HMMs. This resulted in efficient models that could anticipate subject movements and output reasonably accurate predictive sequences that simulate typical subsequent human behavior from initial preparatory gesture analysis.

With the increasing availability of large amounts of annotated datasets and the advent of neural networks, many new models have been introduced for the purpose of behavior prediction, with alarmingly accurate results, thus creating a new paradigm for model-free, highly flexible and non-linear generalization[55][65]. Time-dependent neural networks make use of a time-delay factor and use preceding values for future predictions[66-67]. Self-organizing Kohonen networks could efficiently learn characteristics of normal trajectories and thus detect novel or anomalous behavior [68], comparable to earlier non-linear SVM models[69]. Game theoretic deep learning models and adversarial networks endeavor to simulate multi-agent strategic interactions [70]. More recently, sensor-based recurrent neural network models have been introduced that attempt to remember previously seen activity transitions in

smart home environments and predict typical behavior from given initial state information[71].

Although these models did not explicitly specify activity recognition as an essential prerequisite to behavioral prediction, it has since been agreed upon that identification of the ongoing activity can in fact significantly improve the predictions on future behavioral patterns[72]. This pattern is particularly observable in state-based prediction models. Since our paper focuses on vision-based paradigms, we have restricted our review of action recognition systems to vision-based models. The general framework for action recognition system are similar to behavioral prediction models, with a low level processing that takes care of background subtraction, feature extraction, object detection and tracking, mid-level activity state estimation and a high level reasoning engine that outputs predictions. Approaches to activity recognition are either single-layered (space-time or sequential approaches) or hierarchical (statistical, syntactic or description-based approaches). Single layered space-time paradigms are essentially divided into feature-based (global or local activity representations)[57][59][73] that are commonly used in intelligent surveillance/tracking or smart home systems, or cognitive psychology research - inspired human body-model based (kinematics or 3D pose estimations) [74] that are primarily used for gesture recognition in gaming systems. The most noticeable limitation in the case of the latter class of models is that it is computationally expensive and prone to failure in cases of partial occlusion and multiple simultaneous activities. Our paper is in part inspired by the idea behind initial feature-based exploratory models that tried to capture activity as a collection of space-time interest points [73], but by way of neural networks we have overcome the many non-trivial encumbrances involved in the implementation of these early models, including the selection of optimal space time descriptors and clustering algorithms for minimally sized codebook construction.

Advancements in activity recognition have been most significantly noticeable in recent developments in video description models. These descriptive models use embedded information from previously seen training data like a visual vocabulary to optimize classification of new, unseen videos [75]. By incorporating memory modules[76] and recurrent loops[47] into the system, they have been able to incorporate the semantic and hierarchical nature of complex dynamic scenes and events into predictions. Although effective in appropriately constrained, uncluttered environments, the deterministic nature of these models makes them more susceptible to noise and occlusion that is inevitable in natural environments.

Fig. 2. Sample retrieval of Spatio-Temporally Salient Regions by our model in low resolution videos of concurrent activity.

3 Saliency Based Action Recognition Model

In this section, we present an LSTM-based model for robust and accurate activity recognition, augmented by a pre-trained spatio-temporal saliency detection network for the purpose of implicit activity localization and embedding of contextually significant information found in successive video frames. Human activity recognition is a vital field of interest in computer vision research. It has wide-ranging applications that include intelligent surveillance, patient monitoring systems and a variety of upcoming technology that incorporates human-computer interfaces into everyday devices.

In our model, we use visual attention to identify the most conspicuous distinctive characteristics of an activity followed by bi-directional Long Short Term Memory (LSTM) to remember and recognize these descriptive individualities in previously unseen videos, and output predictive labels classifying the ongoing activities. The Visual Attention module essentially acts as a spatial auto-encoder for perception and motion saliency.

Fig. 3. Sample STSHI output for a window of 30 frames that captures the action of opening a computer.

The Memory module, by taking as input the salient regions in consecutive frames, records temporal information and learns transitory semantics characteristic of each frame, while also retaining contextually relevant information that provides crucial cues about the environment. Theoretically, this approach should be feasible in uncontrolled, cluttered environments as well, since the prominent regularities that enable us to identify the occurrence of a particular activity, should still be perceptible to the visual attention network, and thus comprehensible to the memory module.

The saliency detection network we have used for our visual attention module is inspired by the model presented in this paper[68], and our current implementation of the Visual Attention Module makes use of the filters provided by OpenCV 3's Saliency module to provide a simple, functional end-to-end prototype for this purpose. Ideally, this would be replaced by a fine-tuned multi-stream convolutional neural network for spatio-temporal saliency detection. However, since most state of the art models are trained primarily on human eye fixation data, they work best on high resolution images of uncluttered environments, and their output is limited to a small number of highly salient points in the image, that cannot effectively capture im-

portant contextual information. Figure 2 is an example to show the effectiveness of our module even in very low resolution video frames and poor indoor lighting conditions. The weight matrix of the visual attention unit is then set to non-trainable while training the LSTM network for activity recognition.

Since the input to the visual attention network is in video format while the input to the memory module is in the format of a series of images, it is important to reduce the temporal aspect of movement in the three-dimensional(X, Y, T) input data into a two-dimensional(X, Y) representation and this is done by concatenating a small window of successive saliency predictions along the time axis.

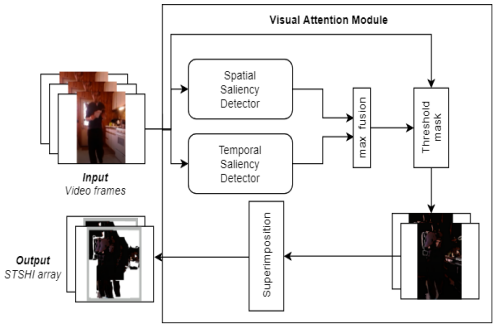


Fig. 4. Overview of the Visual Attention Module.

Superimposition of only the temporally salient regions indicative of motion, extracted using Dense Optic Flow detection, while retaining common spatially salient characteristics, allows us an optimal representation of micro-gestures within a brief temporal window. The input to the memory module is thus similar to Motion History Images[59] or Spatio-Temporal Shape Templates[73], essentially consistent of concatenated blobs of Salient regions in consecutive video frames and shall hereinafter be referred to as Spatio-Temporal Saliency History Images(STSHI) instead. The 3-channel data is finally reduced to a single grayscale thresh map representation. Figure 3 is an example for an STSHI template formed by concatenating features from a window of 30 consecutive video frames, and clearly represents the simple action of a laptop computer being opened. An array of STSHI templates corresponding to a single video is then fed to the bi-directional LSTM network that intrinsically learns a mapping function to match the input STSHI templates to the correct activity category. Figure 4 illustrates the functioning of the Visual Attention Module, and the intermediate representation of the salient regions before superimposition. We have also chosen to maintain the original dimensions of the input frames, so the network may also learn from the relative spatial orientation of salient features, a characteristic from which the mind would immediately derive a gist of the visual scene.

Our memory model was inspired by the idea behind earlier models that, despite having implemented neither visual saliency nor long term dependencies, attempted to ‘match’ videos using spatio-temporal feature relationships[71]. The most recent models of Long Short Term Memory cells[78], have been shown to consistently outperform standard Recurrent Neural Networks in learning tasks that have involved

identification of persistent, long term dependencies. Figure 5 is a simplified depiction of the internal functioning of an LSTM cell. Figure 6 describes the equations for the learning process of an LSTM cell at a given time step *t*.

A recurrent neural network, consistent of two serially connected bi-directional LSTM units has been proposed for our memory module instead of a standard LSTM because this model is intended to be a precursor to a behavior analysis model, and it is useful to simultaneously learn encoding and decoding, from saliency templates to activity labels, and vice versa.

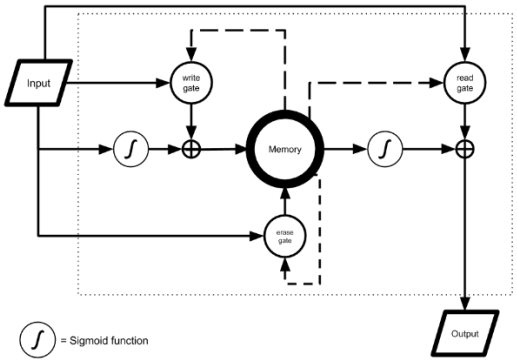


Fig. 5. Components of an LSTM cell

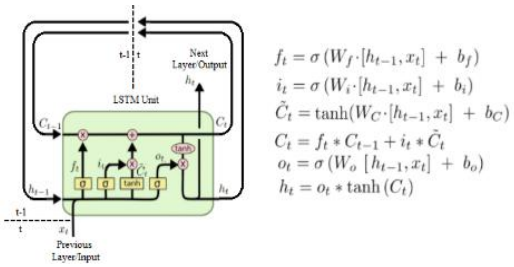


Fig. 6. Learning Process in an LSTM cell

It also allows for some amount of variation in the sequence of primitive gestures that make up an activity and we conjecture that it makes the model invariant to a measurable amount of temporal scaling. During the fully supervised training process, each input to the memory module is essentially the output obtained from the primary module for visual attention and consists of an array of resized STSHI to represent the original input video. Resizing is done in order to reduce the number of features being fed to the network, but care is taken to maintain the original height-to-width ratio so as not to distort the saliency detections. The target output is a binary HOT-encoded vector of length equivalent to total number of possible activity classes, with only the elements corresponding to activities depicted in the video set to 1.

The first bi-directional LSTM unit takes the input and produces one output per template in an STSHI sequence, using the TimeDistributed wrapper layer in Keras.

Since each template essentially captures a micro-gesture in the activity, the first memory unit effectively learns to recognize micro-gestures and then sends a sequence of outputs to the second memory unit. The second bi-directional LSTM unit also utilizes a TimeDistributed wrapper layer, and condenses the series of outputs into a single output vector. This thus reduces the first layer's outputs, which corresponded to each constituent gesture in an activity video, into a single output corresponding to the overall activity being performed in the video. A dense layer is then used to further format the vector into the shape of the expected output vector. The final output of the memory module is a vector of the same length as the target output vector, and each element of the vector is a probability score denoting the likelihood of the corresponding activity being performed in the input video. By setting an optimum threshold limit upon these confidence predictions, we can successfully eliminate the consideration of unlikely activities, while preserving predictions for possibly concurrent activities in the same video. The easiest way to do this would possibly be to consider as the threshold a percentage (eg: 80%) of the highest confidence score outputted, provided that the maximum is high enough to indicate certain presence of the corresponding activity.

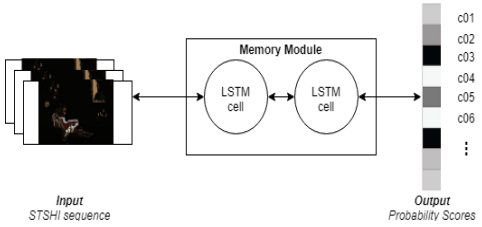


Fig. 7. High-level Overview of the Memory Module

The dataset we will be using for training and testing purposes is the publicly available Charades dataset of daily human activities in indoor environments[79], scaled up to 480p. The videos consist of one or two human subjects performing common activities such as writing, walking, using a computer etc. Figure 7 illustrates the architecture of our Memory Module.

Our entire model can be considered an analogy of the human anatomy responsible for the processing of visual information and therein lies its inherent neurobiological plausibility. The visual attention module's recognition of spatially and temporally significant locations is comparable to the splitting of motion and detail visual information, i.e., the magno and parvo cellular pathways into the cortical area(V1), via the lateral geniculate nucleus(LGN) of the thalamus. Parvocellular ganglion retinal cells being smaller and slower, carry many details such as colour, intensity etc. Magnocellular ganglion retinal cells being larger, faster and rather rough in their representations, carry transient information such as motion details. Since we only retain the most salient regions for the STSHI templates, we also incorporate the aspect of foveated vision at interesting locations. We do not deploy background subtraction as in many popular convolutional models, because activity recognition is highly contextual, and studies have proven that the brain unconsciously utilizes the awareness of several spatial cues to form

illusory conjunctions about the surrounding environment. The memory module models the short term memory of humans and the tendency to recognize the reoccurrence of prior experiences by retaining an internal representation of their distinctive temporal semantics. As such, our model is a better prototype for human activity recognition in real world environments as opposed to activity recognition in controlled virtual simulations or video viewing experiments.

4 Saliency-based Behavioral Prediction – Discussion

Experiments have shown that human behavior is the outcome of interacting determinants and that every individual action is an aggregate of self-intention, social interaction and contextual environment. It is evident from our study that all the developments in saliency detection and behavioral analysis up until now have not completely considered critical factors in the estimation and prediction of a human entity's natural gaze or motion. Saliency detection frameworks have so far failed to account for important aspects of cognitive psychology and behavior that, as prior self-information, are imperative to understanding visual perception in the real world. Behavioral prediction models, on the other hand, have not so far attempted to look into the inevitable influence that first-person perception has upon self-information accumulation and thus, upon any kind of interaction. Here, we conjecture upon a model with the intent to bridge the gap between these two models in the context of intelligent visual surveillance, in order to simulate the ability of a human observer to detect an observable behavioral anomaly in a dynamic visual scene.

As we have shown through our model for visual attention-based activity recognition, detecting spatio-temporally salient regions in a given video can significantly improve predictions about the activity being represented in it. Since our model is a single layered space-time based approach to activity recognition, it is clear that advancing it to incorporate the hierarchical nature of behaviors that are composite of atomic activities, is probably the best possible way to analyze and predict subsequent behavior. Previously, hierarchical models for this purpose adopted symbolic artificial intelligence techniques and were primarily based on statistical (Bayesian Networks [80] and Markov Models [63]) or syntactic (Stochastic Context Free Grammars[81]) approaches, but recent developments in deep learning – specifically recurrent neural networks[71], have significantly outperformed earlier models, and appear to provide a more viable alternative for probabilistic predictive analysis of human behavior. Since all behavioral analysis models ultimately attempt to capture the transitions between internal states within the human brain by way of understanding observable external information, it makes sense that the solution would probably lie within a combination of various neural networks for learning and memory – a concept that is itself based on recreating neuronal interconnections within the brain.

Distinctive behavior could easily be described as individual mannerisms applied to common performative actions. Our proposed model intends to exploit this property by trying to create a relatively change invariant activity recognition model, so as

to be reasonably robust to personal affectations. Thus we can generalize behavioral tendencies based on context and develop a system to predict future actions and detect anomalous or suspicious movements. In future adaptations of our proposed model, we intend to introduce an additional recurrent network architecture, to learn sequences of the activity labels predicted by our current system, and thus classify overall behaviors. This system could then be applied in scenarios where intelligent surveillance of dynamic environments is required, to detect atypical behavior patterns.

References

1. Koch, C. and Ullman, S.: "Shifts in selective visual attention: towards the underlying neural circuitry", Human Neurobiology, 1985.
2. Treisman, A. and Gelade, G.: "A feature-integration theory of attention", Cognitive Psychology, 1980.
3. Itti, L., Koch, C., and Niebur, E.: "A model of saliency-based visual attention for rapid scene analysis", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254–1259, 1998.
4. Ballard, D.H., Hayhoe, M., and Pelz, J.: "Memory representations in natural tasks", J. Cognitive Neuroscience, 7, 66–80, 1995.
5. Ungerleider, S.: "Mechanisms of visual attention in the human cortex", Annual review of neuroscience, 2000.
6. Itti, L. and Koch, C.: "A saliency-based search mechanism for overt and covert shifts of visual attention", Vision Research, 40, 1999.
7. Minut, S. and Mahadevan, S.: "A reinforcement learning model of selective visual attention", Autonomous Agents Conference, 2001.Author, F.: Article title. Journal 2(5), 99–110 (2016).
8. Hillstrom, A.P., and Yantis, S.: "Visual-motion and attentional capture", Perception and Psychophysics, 55, 399–411, 1994.
9. Hamker, F. and Worcester, J.: "Object Detection in Natural Scenes by Feedback", Biologically Motivated Computer Vision: Second International Workshop: 398-407, 2002.
10. Jovancevic, J., Sullivan, B., and Hayhoe, M.: "Control of Attention and Gaze in Complex Environments", Journal of Vision, 1431-50, 2006.
11. Tatler, B.W.: "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions", J. Vis., 11 2007.
12. Hou, X. and Zhang, L.: "Saliency detection: A spectral residual approach", CVPR'07. IEEE, 2007.
13. Goodale, M.A., and Milner, A.D.: "Separate visual pathways for perception and action", Trends in Neurosciences, 15(1):20–25, 1992.
14. Cerf, Moran & Frady, Edward & Koch, Christof: "Faces and text attract gaze independent of the task: Experimental data and computer model", Journal of vision. 9. 10.1-15. 10.1167/9.12.10, 2009.
15. Jovancevic-Misic, J. and Hayhoe, M.: "Adaptive Gaze Control in Natural Environments", Journal of Neuroscience, 29(19), 6234–6238. doi:10.1523/JNEUROSCI.5570-08, 2009.
16. Folk, C.L., Remington, R.W., and Johnston, J.C.: "Involuntary Covert Orienting Is Contingent on Attentional Control Settings", Journal of Experimental Psychology, 18(4): 1030-1044, 1992.

17. Wu, C., Wang, H., and Pomplun, M.: “The roles of scene gist and spatial dependency among objects in the semantic guidance of attention in real-world scenes”, Vision Research, 105: 10-20, 2014.
18. Shinoda, M., Hayhoe, M.M., and Shrivastava, A.: “What controls attention in natural environments”, Vision Research, 41, 2001.
19. Triesch, J., Ballard, D.H., Hayhoe, M.M., and Sullivan, B.T.: “What you see is what you need”, Journal of Vision, vol. 3, 2003.
20. Oliva, A., Torralba, A., Castelhano, M., and Henderson, J.: “Top–down control of visual attention in object detection”, Proceedings of the 2003 International Conference on Image Processing, 2003.
21. Bruce, N. and Tsotsos, J.: “Saliency, attention, and visual search: An information theoretic approach”, JoV, 2009.
22. Lee, T. S., and Stella, X. Y.: “An information-theoretic framework for understanding saccadic eye movements”, NIPS, 1999.
23. Bruce, N. and Tsotsos, J.: “Saliency based on information maximization”, Advances in Neural Information Processing Systems, 2006.
24. Fink, G. R., Dolan, R. J., Halligan, P. W., Marshall, J. C., and Frith, C. D.: “Space-based and object-based visual attention: shared and specific neural domains”, Brain: 2013-28, 1997.
25. Itti, L. and Baldi, P.: “A principled approach to detecting surprising events in video”, Proc. IEEE CVPR, 2005.
26. Judd, T., Ehinger, K., Durand, F., and Torralba, A.: “Learning to predict where humans look*”*, IEEE, 2009.
27. Judd, T., Durand, F. and Torralba, A.: “Fixations on low resolution images*”,* Journal of Vision, 11(4):14–14, 2011.
28. Jiang, M., Huang, S., Duan, J., and Zhao, Q.: “SALICON: Saliency in context”, CVPR 2015, IEEE 2015.
29. Jiang, M., Boix, X., Roig, G., Xu, J., Van Gool, L., and Zhao, Q.: "Learning to predict sequences of human visual fixations", IEEE transactions on neural networks and learning systems, vol. 27, no. 6, pp. 1241-1252, 2016.
30. Cornia, M. , Baraldi, L., Serra, G., and Cucchiara, R.: “Predicting Human Eye Fixations via an LSTM-based Saliency Attention Model*”*, arXiv preprint arXiv:1611.09471, 2017.
31. Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., and Torralba, A.: *MIT Saliency Benchmark*, 2017, [online] Available: http://saliency.mit.edu.
32. Liu, Z. , and Zhang, X. and Luo, S. and Le Meur, O.: “Superpixel-based spatiotemporal saliency detection”, IEEE Trans. on Circuits and Systems for Video Technology, 2014.
33. Kruthiventi, S. S., Gudisa, V., Dholakiya, J. H., and Venkatesh Babu, R.: “Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation”, CPVR, 2016.
34. Dodge, S. and Karam, L.: “Visual Saliency Prediction Using a Mixture of Deep Neural Networks”, arXiv preprint arXiv:1702.00372, 2017.
35. Tatler, B. W., Hayhoe, M. M., Land, M. F., and Ballard, D. H.: “Eye guidance in natural vision: Reinterpreting salience”, Journal of vision, 11(5), 2011.
36. Kümmerer, Matthias et al.: *“*Understanding Low- and High-Level Contributions to Fixation Prediction.” 2017 IEEE International Conference on Computer Vision (ICCV): 4799-4808, 2017.
37. Foulsham, T., Walker, E., and Kingstone, A. : “The where, what and when of gaze allocation in the lab and the natural environment”, Vision Research, 1920-31, 2011.
38. Tatler, B.W. :“Eye movements from laboratory to life*”,* Current Trends in Eye Tracking Research, 17-35, 2014.

39. Kaitlin, E. W. Laidlawa, Foulshamb, T., Kuhnc, G., and Kingstone, A.: "Potential Social Interactions are Important to Social Attention", PNAS, 2011.
40. Gobel, M. S., Kim, H. S., and Richardson, D. C.: "The dual function of social gaze", Cognition: 359-64, 2015.
41. Murabito, F., Spampinato, C., Palazzo, S.. Giordano, D. Pogorelov, K. and Riegler, M.: "Top-down saliency detection driven by visual classification", Computer Vision and Image Understanding, 2018
42. Kruthiventi, S. S., Ayush, K., and Babu, R. V.: "Deepfix: A fully convolutional neural network for predicting human eye fixations", CoRR, abs/1510.02927, 2015.
43. K¨ummerer, M. , Wallis, T.S., Bethge, M.: "Deepgaze II: Reading fixations from deep features trained on object recognition", arXiv preprint arXiv:1610.01563, 2016
44. Cornia, M., Baraldi, L., Serra, G., and Cucchiara, R.: "A deep multilevel network for saliency prediction", Proceedings of the International Conference on Pattern Recognition, 2016.
45. Wang, W., and Shen, J.: "Deep Visual Attention Prediction", arXiv preprint arXiv:1705.02544, 2018.
46. Simonyan, K., and Zisserman, A.: "Two-stream convolutional networks for action recognition in videos", NIPS, pages 568–576, 2014.
47. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T.: "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description", CVPR, 2015.
48. Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., O'Connor, N.E.: "Shallow and deep convolutional networks for saliency prediction", CVPR 2016.
49. Bak, C., Erdem, E., and Erdem, A.: "Two-Stream Convolutional Neural Networks for Dynamic Saliency Prediction", arXiv preprint arXiv arXiv:1607.04730, 2016
50. Bak, C., Kocak, A., Erdem, E., and Erdem, A.: "Spatio-Temporal Networks for Dynamic Saliency Prediction", arXiv preprint arXiv arXiv:1607.04730v2, 2017
51. Kuen, J., Wang, Z. and Wang, G.: "Recurrent Attentional Networks for Saliency Detection", CVPR, 2016.
52. Unzicker, A., Juttner, M. and Rentschler, I.: "Similarity-based models of human visual recognition", Vision Research 38: 2289–2305, 1998.
53. Hu, W., Tan, T., Wang, L., and Maybank, S.: "A Survey on visual surveillance of object motion and behaviors", IEEE Transactions on Systems, Man and Cybernetics, 34(3): 334-352, 2004.
54. Poppe, R.: "A Survey On Vision-based Human Action Recognition", Image and Vision Computing, 28(6) : 976-90, 2010.
55. Van Kasteren, T., Noulas, A., Englebienne, G., and Kröse, B.: "Accurate activity recognition in a home setting", Proceedings of the 10th international conference on Ubiquitous computing, 1-9, 2008
56. Avci, U., and Passerini, A.: "A Fully Unsupervised Approach to Activity Discovery", HBU '13, 77-88, 2013.
57. Bobick, A., Davis, J.: "Real-time recognition of activity using temporal templates", Applications of Computer Vision 1996. WACV '96. Proceedings 3rd IEEE Workshop on, pp. 39-42, 1996.
58. Wilson, A. D., Bobick, A. F., and Cassell, J.: "Temporal classification of natural gesture and application to video coding", Proc. IEEE Conf. Computer Vision and Pattern Recognition, 948–954.,1997.
59. Bobick, A., Davis, J.: "An appearance-based representation of action", ICPR, 1996.
60. Bobick, A. F. and Wilson, A. D.: "A state-based technique to the representation and recognition of gesture", IEEE Trans. Pattern Anal. Machine Intell., 19: 1325–1337, 1997.

61. Rabiner, L.: “A tutorial on hidden Markov models and selected applications in speech recognition”, Proceedings of the IEEE, 1989.
62. Fossler-Lussier, E.: “Markov Models and Hidden Markov Models: A Brief Tutorial”, International Computer Science Insitute, 1998.
63. Pentland, A., and Liu, A.: “Modeling and Prediction of Human Behavior”, Neural Computation 11(1): 229-42,1999.
64. Kim, E., Helal, S., Cook, D.: "Human activity recognition and pattern discovery", IEEE Pervasive Computing, 9 : 48-53, 2010.
65. Phan, N., Dou, D., Piniewski, B., and Kil, D.: “Social Restricted Boltzmann Machine: Human Behavior Prediction in Health Social Networks”, ASONAM '15, 2015.
66. Yang. M. and Ahuja, N.: “Extraction and classification of visual motion pattern recognition”, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 892–897, 1998.
67. Meier, U., Stiefelhagen, R., Yang, J. and Waibel, A.: “Toward unrestricted lip reading”, Int. J. Pattern Recognit. Artificial Intell., 14(5) : 571–585, 2000.
68. Owens, J. and Hunter, A.: “Application of the self-organizing map to trajectory classification”, Proc. IEEE Int. Workshop Visual Surveillance, 2000.
69. Zhao, H. and Liu, Z.: “Human Action Recognition Based on Non-linear SVM Decision Tree”, Journal of Computational Information Systems 7(7): 2461-68, 2011.
70. Hartford, J., Wright, J.R. and Leyton-Brown, K.: “Deep Learning for Human Strategic Behavior Prediction”, NIPS, 2016.
71. Almeida, A. and Azkune, G. “Predicting Human Behavior with Recurrent Neural Networks”, Appl. Sci. ’18, 2018.
72. Sigurdsson, G., Russakovsky, O., and Gupta, A.: “What Actions are Needed for Understanding Human Actions in Videos?”, ICCV ’17, 2156-2165, 2017.
73. Bregonzio, M., Gong, S. and Xiang, T.: "Recognising action as clouds of space-time interest points", Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR), 2009.
74. An, N., Sun, S., Zhao, X., and Hou, Z.: “Remember like humans: Visual tracking with cognitive psychological memory model”, International Journal of Advanced Robotic Systems, 1-9, 2017.
75. Plummer, B.A., Brown, M., and Lazebnik, S.: “Enhancing Video Summarization via Vision-Language Embedding”, CPVR, 2017.
76. Fakoor, R., Mohamed, A., Mitchell, M., Kang, S. B., and Kohli, P.: “Memory-augmented Attention Modelling for Videos”, arXiV preprint, arXiv: 1611.02261v4, 2017.
77. Ryoo, M.S., and Aggarwal, J.K.: “Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities”, Computer vision, 1593–1600, 2009.
78. Hochreiter, S. and Schmidhuber, J.: “Long Short-Term Memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
79. Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., and Gupta, A.: “Hollywood in homes: Crowdsourcing data collection for activity understanding”, ECCV, 2016
80. Oliver, N., Rosario, B., and Pentland, A.: “A Bayesian Computer Vision System for Modeling Human Interactions”, Proceedings of ICVS, 1999.
81. Ryoo, M.S., and Aggarwal, J.K.: “Semantic Representation and Recognition of Continued and Recursive Human Activities”, Int J Comput Vis, 82: 1–24, 2009.