Fernández-Torres, Miguel-Ángel; González-Díaz, Iván; Díaz-de-María, Fernando. (2020). Probabilistic Topic Model for Context-Driven Visual Attention Understanding. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6), pp. 1653-1667.

DOI: 10.1109/TCSVT.2019.2909427

Probabilistic Topic Model for Context-Driven Visual Attention Understanding

Miguel-Ángel Fernández-Torres, Iván González-Díaz, *Member, IEEE*
and Fernando Díaz-de-María, *Member, IEEE.*

***Abstract*—Modern computer vision techniques have to deal with vast amounts of visual data, which implies a computational effort that has often to be accomplished in broad and challenging scenarios. The interest in efficiently solving these image and video applications has led researchers to develop methods to expertly drive the corresponding processing to conspicuous regions that either depend on the context or are based on specific requirements. In this paper, we propose a general hierarchical probabilistic framework, independent of the application scenario, and founded on the most outstanding psychological studies about attention and eye movements which support that guidance is not based directly on the information provided by early visual processes but on a contextual representation arose from them. The approach defines the task of context-driven visual attention as a mixture of latent sub-tasks, which are in turn modeled as a combination of specific distributions associated to low-, mid- and high-level spatio-temporal features. Learning from fixations gathered from human observers, we incorporate an intermediate level between feature extraction and visual attention estimation that enables to obtain comprehensively guiding representations. Experiments show how our proposal successfully learns particularly adapted hierarchical explanations of visual attention in diverse video genres, outperforming several leading models in the literature.**

***Index Terms*—Top-down visual attention, hierarchical probabilistic framework, context-aware model, latent topic models.**

I. Introduction

A great world full of visible information to understand is opened to us, and our visual system has the paramount task of dealing with attentive processes. Due to the limited capacity of the brain to process such a big amount of sensory input, attention involves the inherent search operations that reformulate and optimize generic perception and cognition problems so that they become tractable [1]. Eye movements allow acquiring and tracking visual stimuli, unconsciously highlighting the most conspicuous [2] [3] areas in a particular context, or willingly selecting those that aid to solve a particular task [4].

Computer vision techniques have nowadays to deal with millions and millions of data available, just like the human visual system. This is probably the prime reason why the effort in developing computational systems to accomplish this selective task has increased during the last few years. The purpose of researchers is to address traditional image and video applications, such as object [5] and action [6] recognition, video surveillance [7], video summarization [8] or image quality assessment [9], in broader and more complex scenarios, while providing more efficient solutions and better performances.

Looking from a psychological perspective, two theories have been the most influential for computational attention systems. First, the *Feature Integration Theory* (FIT) [10], introduced by Treisman and Gelade in 1980, stated that several features are identified early, automatically and in parallel across the visual field, while objects are registered separately as a conjunction of these features at a later serial stage. In addition to this theory, Wolfe's *Guided Search Model 2.0* (GSM) [11] in 1994 claimed that this serial search had to be guided by useful information in the parallel processes and not independently, which divided the set of stimulus into distractors and candidate targets. It should also be mentioned the importance of eye movements in scene perception, explained in the famous classic study of Yarbus [12] from 1967. A complete experience of perception is based on both a general abstract representation of the scene and the information provided by fixations.

Based on the foundations of these theories and studies, we can differentiate between two main families of visual attention models: bottom-up and top-down. Bottom-Up (BU) models are mainly based on characteristics of the visual scene (*stimulus-driven*) such as color, orientation, motion or depth. By contrast, Top-Down (TD) models (*goal-driven*) are determined by cognitive phenomena like knowledge, expectations or advanced indications. Despite the great amount of visual attention models developed, most of them are BU approaches, whereas TD architectures are still scarce and very often tailored to well-defined scenarios. In such cases, the evaluation of the whole scheme is performed regardless of the capability of the guidance tool. Besides, few investigations acquire the frequently mentioned concurrence of BU and TD factors. No less important is the lack of use and modeling of spatio-temporal and high-level features to address visual attention in real scenarios or videos.

To overcome these shortcomings, we propose a general hierarchical probabilistic framework to estimate visual attention in videos, which can be applied to different scenarios and tasks by simply learning from human fixations. In our model, TD visual attention is decomposed into mixtures of various latent sub-tasks, which are in turn represented as combinations of low-, mid- and high-level features. Depending on the context, distinct features could draw visual attention. For instance, motion features are useful to follow players and track objects in outdoor scenes, while color, faces or text are

M.A. Fernández-Torres, I. González-Díaz and F. Díaz-de-María are with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Avda. de la Universidad 30, Leganés, Madrid, Spain E-mail: matorres, igonzalez, fdiaz}@tsc.uc3m.es.

Manuscript received April 19, 2005; revised August 26, 2015.

more relevant in TV recordings. However, the fundamental basis of the system is, indeed, generic and independent from the application scenario.

This paper is a continuation of our previous work described in [13], where we first introduced this intermediate level formed by latent sub-tasks that bridges the gap between features and visual attention, and enables to obtain more comprehensive interpretations of guidance. More precisely, this article makes several substantial contributions, which update and extend this work as follows:

1) We generate a categorical binary response for each spatial location to model visual attention, in contrast to the continuous variable used in our previous approach. The system now allows to automatically align the sub-tasks discovered to a binary response by means of a logistic regressor, which fully corresponds to the definition of human fixations.

2) We extend the initial set of basic and novelty spatio-temporal low-level features presented in our first work, including and modeling some new mid- and high-level features related to camera motion estimation and object detection, and taking advantage of powerful paradigms such as the Convolutional Neural Networks (CNNs).

3) We provide an in-depth analysis of our proposal for the first time, giving a meaningful insight about the information reflected in each of the sub-tasks that decompose the visual attention. To this end, we illustrate how our approach successfully learns hierarchical guiding representations adapted to several contexts. Furthermore, we perform a comparison with quite a few methods reported in the literature of visual attention in video.

The remainder of the paper is organized as follows: Section II reviews the most relevant and recent related work in perception and visual attention. Section III presents in detail the generative model proposed and briefly introduces the broad set of features considered in our experiments. Experimental results, together with an analysis of the obtained models and a comparison with *state-of-the-art* methods, are gathered in Section IV. Finally, Section V summarizes our conclusions and motivates and outlines future work.

II. Related Work

A. Psychological basis of visual attention

As mentioned above, both Treisman and Gelade (FIT) [10] and Wolfe (GSM) [11] theories established relevant features for the perception of objects, which form the basis of many of the existing visual attention models. First, FIT [10] and other behavioral analysis mentioned three basic features: intensity or luminance contrast, color and orientation. GSM [11] supported later that attention can be guided towards specific targets by modulating gains associated with low-level features, and enumerated other attributes that humans can appreciate efficiently and thus could be also considered salient in a scene: curvature, texture, scale, size, spatial frequency, motion, shape, luminance onset/offset and depth. Subsequents works by Wolfe [14] [15] introduced the idea of ‘guiding representation’ or guidance as a control device located to one side of the main pathway from early vision to object recognition. It controls the access to the attentional bottleneck, so the guidance is abstracted from the main pathway despite of not being part of the pathway itself. Rather than altering the stimulus such as filters would do, this module guides attention as a CCTV operator working at a public building (e.g. a train station or a university) would do. Based on an abstract representation of some notions (e.g. threat, suspicious object), the operator selects some parts of the scenario to receive more attention than others. Hence, guidance is not based directly on the information provided by early visual processes but on a coarse and contextual representation derived from them. This interpretation of visual attention supports the main assumption of our model and opens the door to an inclusion of an intermediate layer that maps the low level stimuli to an intermediate representation. Furthermore, Wolfe extended the list of attributes that might guide the deployment of attention in [14], raising doubtful cases such as novelty or faces, which are tested as input to our approach, in order to appraise their utility in some contexts.

Moreover, the role of eye movements in scene perception had already been studied before the introduction of perception theories referred above. According to the revision of high-level scene perception research made by J. M. Henderson and A. Hollingworth in [16], it is expected to figure out what are the procedures that control *where* and *how long* each fixation point tends to remain centered at a particular location for a complete understanding of scene perception. Yarbus [12] classic study of 1967 showed that, although first few fixations in a scene seem to be controlled by global characteristics, positions of later fixations are not random but landed on regions that are useful or essential for perception. Eyes are either driven by TD factors that direct fixations toward informative task-driven locations (e.g. cooking, driving) or led to low-level image discontinuities called salient regions (e.g. bright regions, edges). The time the eyes remain in a given region also depends on its visual and semantic properties. The experience of a complete and integrated visual world is thus based on an abstract representation that covers general information about the scene combined with perceptual information arose from fixations. By examining eye movements, it could be possible to infer the underlying factors affecting fixations or task at hand, even to interpret observer thoughts [17]. That is the purpose of our model, which introduces an intermediate level between feature extraction and visual attention computation stages based on the information drawn from fixations. This level consists of latent sub-tasks that can be used to determine why some locations are more conspicuous than others. Thus, rather than directly learn a predictor of human attention over low-level visual features, our method provides a hierarchical interpretation of visual attention, advantageous for further comprehensive analysis.

B. Computational visual attention models

While Koch and Ullman [18] designed a model to combine early vision features, and defined the concept of *Saliency Map* (SM) as a mechanism to model local visual attention driven by the set of visual stimuli in the scene, the first implementation and verification of a BU model, which uses color, intensity and orientation features, was performed by Itti et al. [19]. Harel et al. [20] proposed a saliency algorithm based on graphs, which extracted the same features at different scales. These two representations are the most frequently employed in the literature due to their good performance in a variety of situations. For further information, a wider survey on visual attention modeling is presented by Borji et al. in [21].

Despite their importance in the process of driving visual attention, we still lack of generic TD architectures. TD methods are still limited, and mostly integrated within systems conformed to specific scenarios, being the behavior of these approaches often evaluated at application level. Moreover, few investigations consider the cooperative relationship between BU and TD mechanisms that is advocated by the prevalent studies about attention. Most TD approaches guide attention towards specific targets by modulating gains associated with low-level stimuli. Sprague and Ballard [22] proposed a reinforcement learning method that combines action selection and visual perception in a sidewalk navigation task. Navalpakkam and Itti [23] optimized the integration of BU cues for target detection by maximizing the signal-to-noise ratio of the target vs. background. Peters and Itti [24] computed a task-dependent map based on scene gist and gaze in a video games scenario. Judd et al. [25] trained a linear SVM taking some image features and human fixations to define salient locations. Elazary and Itti [26] proposed a more flexible model that can concurrently select the best features to guide attention and adjust the width of feature detectors. However, in contrast to all these previous attempts, the definition of our model is general and independent of the application or scenario, and may therefore be easily adapted to any scenario of application.

On the other hand, bayesian models are characterized by their capacity to learn from data, taking advantage of data statistics to model the underlying attention process and allowing to obtain interpretable relationships between data and visual fixations. Zhang et al. presented in [27] a probabilistic model that defines saliency as the pointwise mutual information between BU local features and TD search target features. Li et al. [28] proposed a multi-task learning approximation for visual attention in video, where different ranking functions for fusing BU and TD maps were learned depending on the scene content. Our design, instead, models visual attention at each spatial location as a logistic regression over the learned intermediate sub-tasks rather than over the features themselves.

CNNs, the current dominant paradigm for many supervised tasks in computer vision, have been also tested for visual attention achieving promising results, mainly in the still image domain. Among the first attempts to rely on deep learning for saliency estimation, it should be mentioned the use of convnet layers as feature maps carried out by Vig et al. in [29], and the SALICON fine- and coarse-scale model [30]. The latter introduced a large-scale dataset for training new models, annotated by means of a mouse-tracking procedure. These supervised schemes involve training end-to-end models according to a loss function, and unify feature extraction, fusion and saliency prediction in a single structure. This makes more challenging the analysis of these stages, due to the abstract nature of representations at the deepest layers of these strategies. Moreover, it has been reported that they still miss some key elements [31], mostly related to misdetections of people, actions and text, and the relative importance assigned to them when they take place simultaneously. Hence, although their capability of discovering discriminant high-level visual features is out of any doubt, it is therefore necessary to clarify the relationship between the feature maps derived from CNNs and the psychophysical stimuli that guide attention. This implies the development of complementary modules able to provide this mapping, such as the hierarchical method presented in this article, which facilitates the integration with such neural network schemes. Indeed, our intermediate sub-task level can be placed straightforwardly over the top layers of a deep network, as shown in the conducted experiments. To do this, we make use of the features derived from a recently released deep contrast network for salient object detection with pixel-level accuracy [32]. It should finally be pointed out that only a few works have drawn on deep learning to tackle the estimation of visual attention in videos, finetuning models on the optical flow estimated from static images [33], studying car driving-related attentional mechanisms [34] and recognizing human activities [35].

III. Visual Attention Topic Model

In this section we describe in detail the system proposed for visual attention in video, which we have called *visual Attention TOpic Model (ATOM)*, where a set of features, such as the ones introduced in the previous section, is used to learn several related sub-tasks. These sub-tasks automatically lead the attention of the system to the most appealing areas of a scene.

A. Model overview

Our generative model is supported by the following assumption: *Task- or context-driven visual attention in video can be modeled as a mixture of several sub-tasks which, in turn, can be represented as combinations of low-, mid- and high-level spatio-temporal features obtained from video frames*. Depending on the scenario, visual attention may be attracted by several different events. Our goal is not to detect these events of interest for a particular application, but to efficiently guide the later processing to areas and time segments of special importance in the video.

Figure 1 illustrates our hypothesis for three different scenarios in CRCNS-ORIG [36] database. First, looking at the contexts given, visual attention may be attracted by different events or elements in the scene: people *running* and *walking* in the case of *Outdoor*; *game character* and *goals* or *items* for *Videogames*; and *players* and *scoreboards* in *Sports*. Note that some contexts may share similar attractions, like *ball*, which is present both on *Outdoor* and *Sports* videos. Our goal is to automatically discover sub-tasks that guide later processing to the areas where those occur, with the purpose of making it simpler. In turn, these sub-tasks can be modeled as combinations of spatio-temporal features. For instance, the use of a motion feature combined with a face or pedestrians

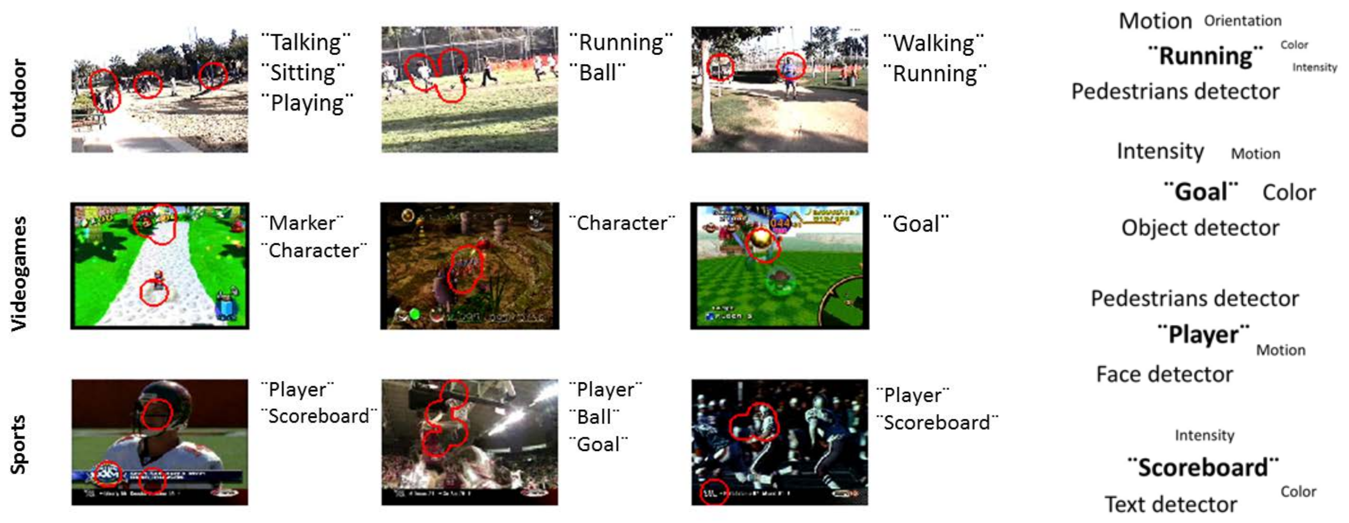


Fig. 1. Visual attention modeled in three different scenarios taken from CRCNS-ORIG [36] database (*Outdoor*, *Videogames* and *Sports*) as a mixture of several relevant sub-tasks (e.g. "Running", "Goal", "Player", etc.), associated with particular areas of special importance for observers, which are highlighted in the example frames on the left side. Some of them may appear similarly in different contexts, such as "Ball" or "Goal". On the right side of the figure, the word clouds show how some sub-tasks (bold central words) are represented as a combination of features (surrounding words: intensity, motion, detectors, etc.). Feature importance, represented by the font size of each word showing a feature, varies from one sub-task to another. For example, motion information and pedestrian detections are more relevant for "Running"; in contrast, an object detector, along with intensity and color features are more advantageous to represent a "Goal" in a videogame.

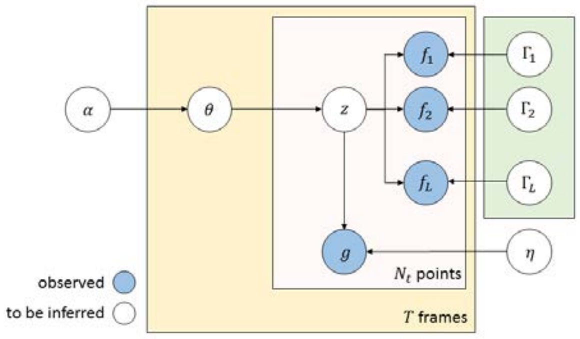


Fig. 2. Graphical representation of the proposed generative topic model for visual attention. Shaded circles represent observations from frames, white circles indicate hidden variables to be inferred, and boxes mean independent repetitions.

detector could be useful to represent "Player" sub-task. In contrast, a "Scoreboard" is well-defined by some intensity or color features, together with a text detector.

Probabilistic Latent Topic Models (LTM), which have been commonly used to extract hidden semantic structures (*topics*) from text corpus, can be helpful to unsupervisely understand large amounts of information, such as the human perception features that are quickly and parallely processed by the brain. Our approach involves thus a LTM which relies on the well-known *Latent Dirichlet Allocation* (LDA) algorithm [37] and some of its supervised extensions [38] [39].

First, by understanding frames as a mixture over topics, LDA enables to interpret how they are composed in a statistical and unsupervised way, associating each frame to multiple topics with different proportions. In our particular scenario, task-driven visual attention is modeled as a finite mixture over a set of K topics, which represent the sub-tasks contributing to model visual attention, either by attracting or by inhibiting it. Note that both terms, topics and sub-tasks, are used interchangeably in the article. In parallel, for a given video frame I_t, a set of L visual descriptors $\mathbf{f} = \{f_1, f_2, ..., f_L\}$ is computed at each spatial location n, so that the latent topics are in turn modeled as combinations of these features.

The original LDA is completely unsupervised, so that the topics are learned to maximize the likelihood of a corpus, and require of human knowledge for an intelligible analysis. In our case, in contrast, we aim to learn how humans guide their attention to visual stimuli and the ground-truth (GT) fixations provided by different subjects will drive our training step. Visual attention is thus estimated by means of a regression model over the topic assignments.

In our previous approach [13], *Supervised Latent Dirichlet Allocation* (sLDA) was used to predict a continuous response variable g_n (the visual attention) for each spatial location $n \in N_t$ in a frame I_t. In this work, we decide to replace this linear regressor by a logistic one, in order to automatically align the topics discovered from frames to the information gathered in GT binary fixation maps. This model draws on the *Dirichlet-Bernoulli Alignment* (DBA) introduced in [39]. Let us note that the latent nature of the topics remains unchanged in our supervised model, as the human fixations used in the training phase are not supervising the topics but, instead, an additional binary response variable which is computed as a logistic regression over the topic proportions.

The proposed ATOM involves the following generative process for each frame I_t in a video corpus $\mathcal{I} = \{I_1, I_2, ..., I_T\}$. Let us note that, for simplicity, we have removed the sub-index t of the frame in the notation:

1) Draw the frame particular proportions θ of K topics using a corpus-level Dirichlet distribution of parameter α: $\theta|\alpha \sim Dir(\alpha)$.
2) For each spatial location $n \in N$ in a frame I:
    a) Draw topic assignment using a multinomial distribution over the topic proportions θ: $\mathbf{z_n}|\theta \sim Mult(\theta)$.
    b) Represent the local appearance of the spatial location n by drawing L visual features using the topic particular distributions $p(f_{ln}|\mathbf{z_n}, \Gamma_{\mathbf{z_n}l})$, where $\Gamma_{\mathbf{z_n}l}$ are the parameters of the distributions given the selected topic $\mathbf{z_n}$.
    c) Draw the binary response variable g_n modeling the visual attention using a logistic regression model given by the following Bernoulli distribution: $g_n|\mathbf{z_n}, \eta \sim Be(\frac{\exp{(\eta^T \mathbf{z_n})}}{1+\exp{(\eta^T \mathbf{z_n})}})$, where η is the parameter vector that models attention based on the selected topic $\mathbf{z_n}$.

A graphical representation of the model is shown in Figure 2. Intuitively, the K latent topics represent the sub-tasks that contribute to model visual attention. Parameter α represents the prior distribution over the sub-tasks, so that it gives an intuition about the confidence in them. Let us note that some of these sub-tasks may attract human attention whereas others may inhibit it. High values of α_k result in mixtures where all sub-tasks are considered to estimate visual attention in every video frame. In contrast, low values of α_k provide more particular mixtures of sub-tasks for each frame, being the attention determined by only few prevailing sub-tasks. Hence, for each frame I_t, we first generate a particular mixture of these topics θ based on the distribution with the global topic proportions α. Once θ is known, we analyze the different spatial locations of the frame such that, for each n, we first select a sub-task by using the index-variable $\mathbf{z_n}$ ($\mathbf{z_n}$ is an indexing K-dimensional vector with all zeros in except of a 1 in the position of the selected topic). Based on $\mathbf{z_n}$, we draw the local appearance of the spatial location using the particular feature-topic distribution $f_{nl}|\mathbf{z_n}, \Gamma_{\mathbf{z_n}l}$, where $\Gamma_{\mathbf{z_n}l}$ stands for the parameters of the distribution. Sub-task is thus chosen so that its corresponding distribution parameters are the ones that maximize the likelihood of the visual features observed at this location. For the sake of simplicity, we assume that $p(\mathbf{z_n}|\theta)$ is independent for all locations n, which makes the solution tractable, both simplifying the definition of the algorithm and, at the same time, improving the system efficiency. In contrast, other approaches such as Markov Random Fields (MRF) [40], applied in image segmentation, are able to capture such spatial constrains. Nonetheless, it should be noted that some of the visual features that we extract for each sampled location (e.g. color, intensity, orientation, CNNs-based) consider beforehand this spatial dependency. Moreover, we assume conditional independence among features, so that the joint distribution of features for a particular topic can be factorized into the individual probability distributions $p(f_l|\mathbf{z}, \mathbf{\Gamma}_1)$. Finally, we also generate the attention response g_n by computing the logistic regression model over the selected topics.

In contrast, during variational inference, we work on expected values. This means that the indexing variable $\mathbf{z_n}$ is replaced by the variational ϕ_n, which now contains the expected values of the topic assignments given a location n. Therefore, since ϕ_n is a vector with real values (the topic proportions for that sampled location), in practice each location n is in turn modeled as the mixture of sub-tasks that best explains its visual appearance.

B. Guiding features extraction

According to the most leading psychology theories for computational attention systems [10] [11], different simple features are early and pre-attentively processed in parallel to guide visual search in the human brain. Selective visual attention is built on what it is called the *early representation*, a set of conspicuity maps related with some *elementary features* such as color, orientation or motion. These topographical maps do not only surround physical attributes, but also may be explained as relational aspects of these physical characteristics. We may even guide our attention by focusing on mid- and high-level features such as symmetry, faces or text.

Motivated by the general conclusions of these theories, ATOM may operate over a great number of diverse features. Depending on their nature, they may be modeled using various probability distributions: e.g. *normal*, *exponential*, *discrete*, etc. It should be remarked that our model is not feature-dependent, so that any kind of feature can be incorporated by selecting the appropriate distribution. Furthermore, for each application scenario and based on human fixations, our model will automatically discover which particular features are more and less discriminant to model attention and correspondingly assign appropriate parameters to their distributions. Hence, one could include a broad general set of features as the model will automatically diminish the influence of those that do not guide the attention in a particular context. In the experiments contained in this paper, a total set of 18 features has been considered. For the sake of completeness, we briefly describe the features and their corresponding distributions in the following sections. Some of the features are handcrafted and allow us to perform a meaningful interpretation of the estimated visual attention; other, such as those derived from a CNNs are less interpretable but help to improve the system performance.

1) Basic and novelty spatio-temporal features

Firstly, we make use of the conspicuity maps provided by Itti et al. in [19] to consider three commonly-used early visual features: *color* (C), *intensity contrast* (I) and *orientation* (O). Then, the optical flow method introduced in [41] is applied to obtain motion vectors for each spatial location n in a given frame. After, we compute two maps based on them: *velocity* (M) or motion magnitude M_{n_t} (calculated using the L2-norm), and *acceleration* (A), which is its absolute derivative $A_{n_t} = |M_{n_t} - M_{n_t \ 1}|$.

Moreover, those regions of the scene that continually change may also attract the attention of observers. In order to highlight them, *novelty* is modeled by handling *coherence-based*

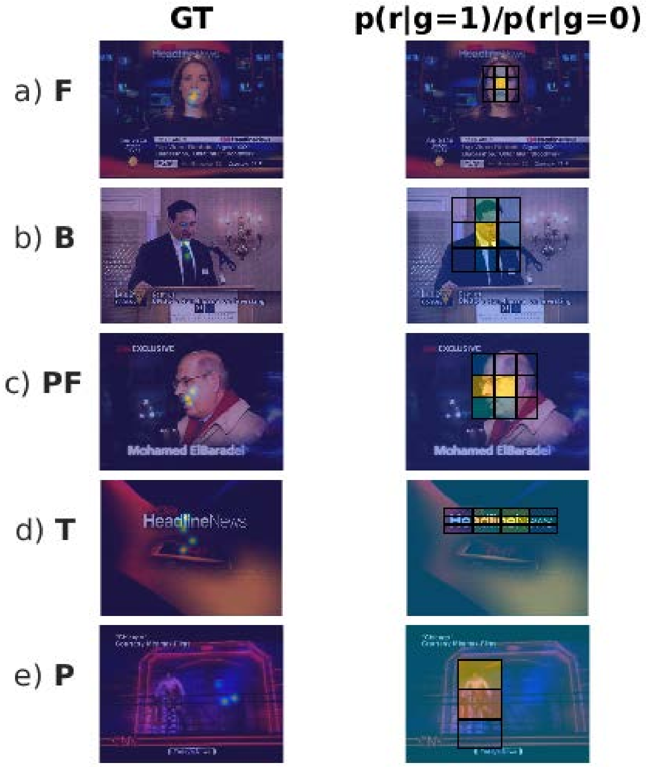


Fig. 3. Object-based feature maps computed for example frames taken from TVNews (a, b, c, d) and TalkShows (e) categories from CRCNS-ORIG [36] database. (Left) Human fixations do not cover the whole object, but concentrate on particular areas/parts of the objects. (Right) Consequently, and based on the detected bounding box, we have divided the image into a set of subregions $r = 0...R$. Some of them ($r > 0$) divide the object into several cells (9 for frontal (F) and profile faces (PF), and upper bodies (B); 3 for pedestrians (P) and 12 for text (T)). Moreover, an additional subregion $r = 0$ is considered for the background, covering the rest of the image. Overlay heat maps highlight subregions where probabilities of each object for fixated points ($p(r|g = 1)$, being $g \in \{0, 1\}$ the ground truth variable indicating if the spatial location attracts or inhibits the attention) are substantially higher than those for non-fixated points ($p(r|g = 0)$). Although the prior probability of objects is fairly lower than the probability of background in the database, it can be seen that objects are quite attractive for observers, due to the significant probability of internal cells given fixated locations.

features, which analyze the distribution of pixel values along space and time to detect areas where dispersion is large. To do this, we rely on the work done in [42], extracting *spatial* (SC), *temporal* (TC) and *spatio-temporal* (STC) coherence maps. For the sake of simplicity, we take the *variance* as scattering measure for all maps. In total, 6 maps are computed: three over the pixel intensity values and three over the motion phase θ_{M_n}.

All these features, which carry continuous values, are modeled using a Gaussian probability density function.

2) Camera motion modeling

Camera motion may also influence viewers regarding a video. Indeed, as seeing in previous studies [43], observers tend to follow the camera motion direction to draw their attention to the new information and objects that emerge in the camera view.

First, let us introduce the notation: $\mathbf{x}_n$ is a 2D vector with spatial coordinates x and y associated to the spatial location n used along the paper. Hence, the visual attention based on camera motion is modeled by means of a 2D Normal distribution over the spatial coordinates $N(\mathbf{c}_{z_n} \odot \mathbf{u}, \Sigma_{z_n})$, where $\mathbf{u}$ is the vector modeling the camera motion as a simple translation whose values are computed from a parametric similarity motion model; $\odot$ stands for the Hadamard product between vectors, and $\mathbf{c}_{z_n}$ is the vector of parameters that establishes a relation between the camera motion and the predicted position of the attention, and is learned during the training process. The second parameter Σ_{z_n}, which controls the spatial extent of the Gaussian distribution, has been empirically set to $\Sigma_{z_n} = diag(0.25)$ in order to cover a sufficiently wide area in the scene.

3) Object-based features

In our experiments, we have included detectors for some general-purpose objects that tend to attract visual attention. In particular, cascade object detectors based on the Viola-Jones algorithm [44] are used to detect people's *frontal* (F) and *profile faces* (PF), *upper bodies* (B) and *pedestrians* (P), and a detector working on the Harris corner response [45] is used to detect *text* (T). However, many detectors for other visual concepts may also be included in our model without effort.

We use the output of these detectors (bounding boxes) to generate high level spatial feature maps. Visual attention usually points to particular locations within the objects, so this fact has to be considered when modeling these features. Since the size of the detected bounding boxes is often large, if we use a 2D Gaussian centered in the bounding box that contains a particular object, for instance, we are notably emphasizing the center of the object with respect to its surroundings. However, attention may be fixed at some elements of the object and not only at its center, such as in the case of faces or pedestrians, where subjects often look at the eyes or upper body part, respectively. Rather than directly considering the detected boxes as the feature maps, we have used them to generate more intelligent discrete spatial distributions. As shown in Figure 3, given a detected bounding box, we consider a non-uniform grid with R+1 cells: R cells ($r > 0$) subdivide the detected box into r small subregions, and an additional subregion is considered for the background ($r = 0$). Hence, for a given object l being detected (we keep l as the index of the features, in this case object detections), we model a discrete distribution over the R+1 defined cells as $p(r|\mathbf{z_n}, \beta_{lz_n})$, where r is a cell in the grid (which is object dependent), and $\beta_{l\mathbf{z_n}}$ are the parameters of the discrete distribution for the object l and the topic $\mathbf{z_n}$. The distributions are then factorized for every object category and instance (in case that more than one object of a given category are detected on the same frame). By means of discrete spatial distributions that divide objects in several sub-regions, we are able to learn which parts of the object are more attractive, taking advantage of this knowledge to provide more accurate estimations of visual attention.

4) CNNs-based features

Finally, we make use of 6 features derived from a CNN for salient object detection. The reason is twofold: first, they allow modeling more general objects than those identified by previously mentioned detectors; and second, they demonstrates the ability of our model to find efficient and diverse combinations of features that help to understand how visual attention works in a given scenario. Features have been drawn from the Deep Contrast Network recently introduced by Li et al. in [32]. We employ the models trained by the authors on a different image dataset, and use the feature maps of the penultimate layer to obtain features modeling general objectness. These feature maps $f_l \in [0,1], l = \{1...6\}$ are then modeled using Gaussian distributions, as we did with the aforementioned basic and novelty spatio-temporal features.

C. Inference

This section explains the inference process of our probabilistic model. As in the original LDA [37] and its extensions [38] [39], exact inference is not possible due to the coupling between the variables θ and $\mathbf{z}$, which prevents from inferring the posterior distribution of the parameters given the data. Therefore, we propose to use a simplified variational distribution q (that is tractable) and mean-field variational inference, so that the Kullback-Leibler divergence between the variational distribution q and the posterior distribution is computed. The proposed variational distribution is as follows:

$$q(\theta, \mathbf{z}|\gamma, \phi_{1:N}) = q(\theta|\gamma)\prod_{n=1}^{N} q(\mathbf{z_n}|\phi_n) \tag{1}$$

that incorporates two new variational parameters: ϕ, which is the parameter of a multinomial distribution $q(\mathbf{z_n}|\boldsymbol{\phi_n})$, and γ, the parameter of a Dirichlet distribution $q(\theta|\gamma)$. This optimization is equivalent to maximize the lower bound (ELBO) over the log-likelihood of all the frames in the corpus. In particular, using Jensen's inequality, the ELBO of the log-likelihood of a frame can be expressed as:

$$\begin{aligned}&log\ p(f_{1:N,1:L}, g_{1:N}|\alpha, \Gamma_{1:K,1:L}, \eta) \geq E_q[log\ p(\theta|\alpha)]\\&+\sum_{n=1}^{N} E_q[log\ p(\mathbf{z_n}|\theta)] + \sum_{n=1}^{N} E_q[log\ p(f_{n,1:L}|\mathbf{z_n}, \Gamma_{1:K,1:L})]\\&+\sum_{n=1}^{N} E_q[log\ p(g_n|\mathbf{z_n}, \eta)] + H(q)\end{aligned} \tag{2}$$

where $E_q[\cdot]$ and $H(\cdot)$ are, respectively, the expectation over the variational distribution q and the entropy of a distribution.

The first two terms of Eq. (2) and the entropy of the variational distribution are identical to the corresponding terms in the ELBO for unsupervised LDA and are described in [37]. The third term is the expected log probability of the data given the related topic model parameters. As was mentioned in Section III-A, we assume conditional independence among features. In the following paragraphs, we particularize this expression for the considered distributions.

If the feature map f_{nl} is modeled with a univariate *Gaussian distribution* then, $\Gamma_{1:K,l} \sim \{\mu_{1:K,l}, \sigma^2_{1:K,l}\}$ such as for basic and novelty spatio-temporal features or CNN-based features, the equation for this term is:

$$\begin{aligned}E_q[log\ p(f_{nl}|\mathbf{z_n}, \Gamma_{1:K,l})] &= -\sum_{k=1}^{K}\phi_{nk}\log(\sigma_{kl}\sqrt{2\pi})\\-\sum_{k=1}^{K}\phi_{nk}\frac{(f_{nl}-\mu_{kl})^2}{2\sigma^2_{kl}}\end{aligned} \tag{3}$$

where ϕ_{nk} is the probability that the location n has been drawn by the topic k.

In the case of camera motion features, the distribution is a multivariate Gaussian $p(\mathbf{x}_n|\mathbf{z_n}, \mu_k, \Sigma_k)$ with $\mu_k = \mathbf{c}_k \odot \mathbf{u}$. However, due to the diagonal nature of the covariance matrix Σ_k we can decompose it into two independent univariate Gaussians and apply the previous expression.

In contrast, if the feature is modeled as a *discrete probability distribution* over cells r in a grid, as happens for objects-based features, the expression is:

$$E_q[log\ p(r_n|\mathbf{z_n}, \beta_{lz_n})] = \sum_{k=1}^{K}\phi_{nk}\log(\beta_{klr_n}) \tag{4}$$

where r_n stands for the region in the non-uniform grid defined for the object l that contains the location n, β_{klr_n} is the value of the of the discrete distribution in that region for the object l and the topic k.

The fourth term includes the visual attention binary response variable g_n and is drawn as a logistic regression model over the topic assignment $\mathbf{z_n}$ with parameter η:

$$\begin{aligned}E_q[log\ p(g_n|\mathbf{z_n}, \eta)] &= E_q\left[\left(g_n - \frac{1}{2}\right)\eta^T\mathbf{z_n}\right]\\-E_q&\left[log\left(exp\left(\frac{\eta^T\mathbf{z_n}}{2}\right) + exp\left(\frac{-\eta^T\mathbf{z_n}}{2}\right)\right)\right]\end{aligned} \tag{5}$$

where $\mathbf{z_n}$ is the vector of topic proportions z_{nk} in the location n. By taking second derivatives, it can be noticed that the second term above is a convex function in the variable $\eta^{T^2}\mathbf{z_n^2} = (\boldsymbol{\eta^T} \odot \boldsymbol{\eta^T})(\mathbf{z_n} \odot \mathbf{z_n})$, so we can bound it by using the lower bound for logistic function, which is the first order Taylor expansion in the variable $\eta^{T^2}\mathbf{z_n^2}$:

$$\begin{aligned}&log\left(exp\left(\frac{\eta^T\mathbf{z_n}}{2}\right) + exp\left(\frac{-\eta^T\mathbf{z_n}}{2}\right)\right)\\&\geq -\frac{\xi_n}{2} - log(1 + exp(-\xi_n))\\&-\frac{1}{4\xi_n}tanh\left(\frac{\xi_n}{2}\right)E_q\left[\eta^{T^2}\mathbf{z_n^2} - \xi_n^2\right]\\&\approx -\frac{\xi_n}{2} - log(1 + exp(-\xi_n))\\&-\frac{1}{4\xi_n}tanh\left(\frac{\xi_n}{2}\right)(\eta^{T^2}\phi_n - \xi_n^2)\end{aligned} \tag{6}$$

where ξ_n is an additional variational parameter associated to each point n.

Computing the derivatives of the KL divergence with respect to the parameters and setting them equal to zero allows us to obtain the update equations for the variational procedure.

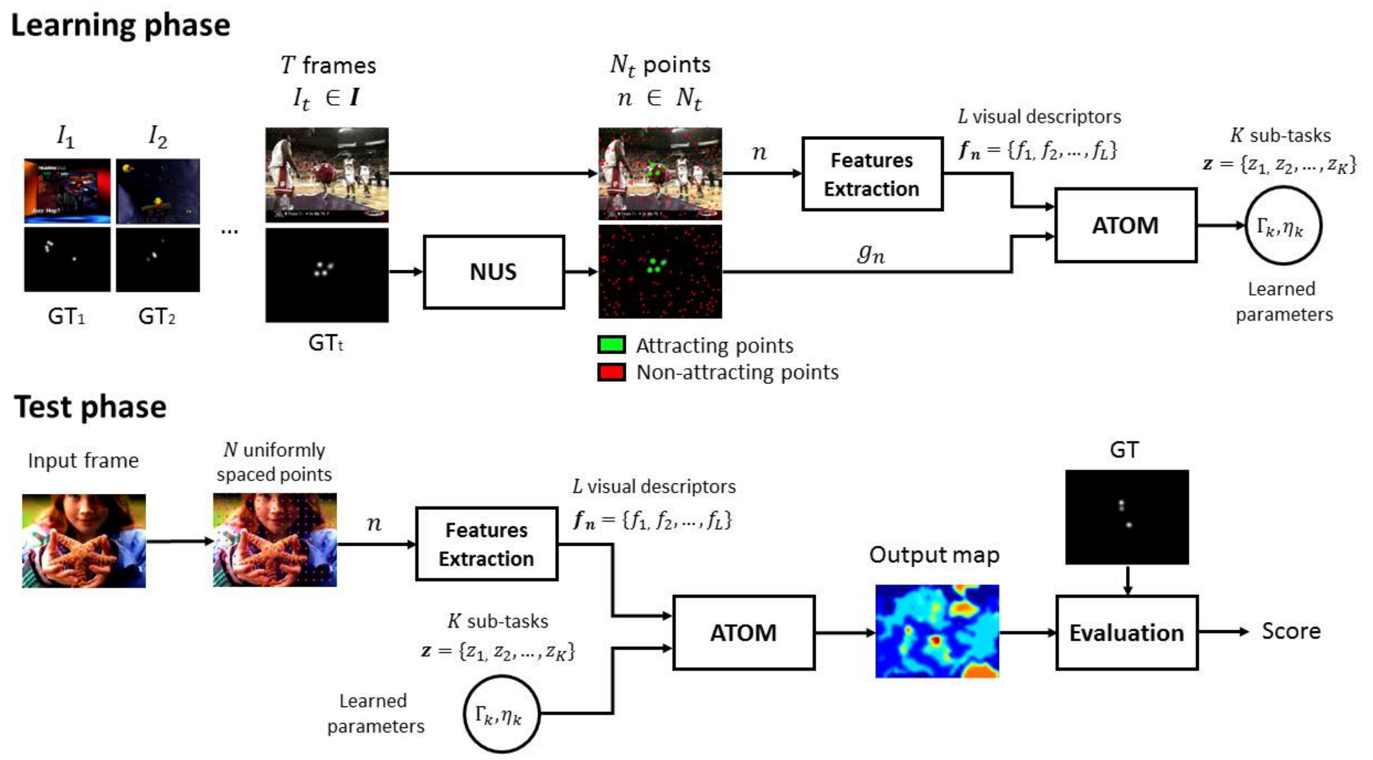


Fig. 4. Processing pipelines of the proposed approach. First, in the learning phase, we learn the optimal values for the parameters associated to the K sub-tasks that model visual attention. A *Non-uniform Sampling (NUS)* strategy allows to generate training datasets that balance the number of attracting and non-attracting points. Then, in the test phase, attention is predicted for each frame at N uniformly spaced locations.

In particular, in the *variational E-step* we must update the variational parameters:

$$\begin{aligned}\phi_{nk} \propto & \frac{\prod_{l=1}^{L_D} \beta_{klr_n}}{\prod_{l=1}^{L_C} \sigma_{kl}} \exp\left[\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right) + \right. \\ & \left(g_n - \frac{1}{2}\right)\eta_k - \frac{1}{4\xi_k} tanh\left(\frac{\xi_k}{2}\right)\eta_k^2 - \\ & \left. \sum_{l=1}^{L_C} \frac{(f_{nl} - \mu_{kl})^2}{2\sigma_{kl}^2}\right] \end{aligned} \tag{7}$$

$$\gamma_k = \alpha_k + \sum_{n=1}^{N} \phi_{nk} \tag{8}$$

$$\xi_{nk} = \eta_k \phi_{nk} \tag{9}$$

being L_C and L_D the number of continuous (Gaussian) and discrete features respectively, and $L = L_C + L_D$ the total number of features. Note that we have used the expression $E_q[log(p(\theta_k|\gamma)] = \Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)$, where $\Psi(\cdot)$ is the digamma function.

In the M-step, we maximize the corpus-level ELBO with respect to the model parameters $\Gamma_{1:K,1:L}, \eta$, in order to compute their optimal values.

First, parameters μ_{kl} and σ_{kl}^2 are computed for each Gaussian feature l and topic k.

$$\mu_{kl} = \frac{1}{\Delta_{kl}} \sum_{t=1}^{T} \sum_{n=1}^{N_t} \phi_{tnk} f_{tnkl} \tag{10}$$

$$\sigma_{kl}^2 = \frac{1}{\Delta_{kl}} \sum_{t=1}^{T} \sum_{n=1}^{N_t} \phi_{tnk} (f_{tnkl} - \mu_{kl})^2 \tag{11}$$

where $\Delta_{kl} = \sum_{t=1}^{T} \sum_{n=1}^{N_t} \phi_{tnk}$ is the normalization factor.

In the case of camera motion, as mentioned above, the parameter is the vector $\mathbf{c_k} = (c_{kx}, c_{ky})$ that multiplies the camera motion vector $\mathbf{u} = (u_x, u_y)$ to determine the mean of the Gaussian distribution:

$$\mathbf{c_k} = \frac{\sum_{t=1}^{T} \sum_{n=1}^{N_t} \phi_{tnk} \mathbf{u_t} \mathbf{x_n}}{\sum_{t=1}^{T} \sum_{n=1}^{N_t} \phi_{tnk} \mathbf{u_t}^2} \tag{12}$$

where $\mathbf{x_n} = (x_{nx}, x_{ny})$ stands for the spatial coordinates vector of the location n.

Finally, for the case of object-based discrete features, the probabilities β_{klr} of the regions r defined on the object-detector l and for every topic k are:

$$\beta_{klr} \propto \sum_{t=1}^{T} \sum_{n=1}^{N_t} \phi_{tnk} 1[r_{nl} = r] \tag{13}$$

where $1[r_{nl} = r]$ means that we have a 1 just in case the region of the point n for the detector l is r (otherwise we have a zero). It is worth noting that we have added the subindex t when necessary to indicate the frame number in the corpus.

Furthermore, during the training step, we use the GT response value g_{tn} of all points in the corpus to learn the parameter of the logistic regression model:

$$\eta_k = \frac{2\sum_{t=1}^{T}\sum_{n=1}^{N_t}\phi_{tnk}(g_{tn}-\frac{1}{2})}{\sum_{t=1}^{T}\sum_{n=1}^{N_t}\frac{\phi_{tnk}}{\xi_{nk}}tanh(\frac{\xi_{nk}}{2})} \tag{14}$$

D. Learning sub-tasks for visual attention estimation

As in other supervised approaches, we can distinguish two main stages in our framework, as shown in Figure 4. First, in the learning phase, optimal values for the parameters that maximize the ELBO of the log-likelihood are learned. As we need to learn from annotated data, we first describe how we sample this data from the annotated video datasets. Since we are on a highly unbalanced scenario, in which the areas that attract visual attention are strongly less prominent that those that inhibit it, we need to prevent the later dominating the learning process, which might lead to a poor performance. For that end, we have used the *Non-uniform Sampling (NUS)* strategy proposed in [46], which allows to generate training datasets that balance the number of attracting and non-attracting points. While the first are selected based on the GT masks computed from human fixations for a given video frame, non-attracting points are sampled from those spatial locations which have not been fixated by viewers in any frame of the same video. In addition, the sampling process also provides the ground truth binary response g_n for each sampled spatial location ($g_n = 1$ for attracting points, and zero otherwise).

Once models are trained, in the test phase, attention is predicted at uniformly spaced locations n in frames. For that end, we remove all terms relating to the supervision (variable g) and estimate the visual attention maps using the expected value of the logistic regression over the topic assignments:

$$E[g_n|f_{n,1:L},\alpha,\Gamma_{1:K},\eta] \approx \frac{\exp{(\eta^T\phi_n)}}{1+\exp{(\eta^T\phi_n)}} \tag{15}$$

In addition, knowing that given a particular frame visual attention is usually focused on small areas of the size occupied by fixations, a histogram equalization procedure is carried out to highlight the most significant regions detected, which helps to improve the system performance.

IV. Experiments

A. Experimental design

Database

The purpose of our experiments is to demonstrate the ability of the proposed ATOM to learn meaningful sub-tasks that can be used to understand what guides visual attention in different contexts, drawing conclusions on whether observers are either driven by similar generic sub-tasks or, in contrast, by certain specific tasks related to each particular scenario. For this reason, we have selected the well-known freely-accessible CRCNS-ORIG [36] as benchmark dataset. The database contains eye movement recordings from eight distinct subjects freely watching 50 different video clips (over 46,000 video frames, 25 minutes total, 640 × 480). Eye traces have been obtained using a 240 Hz ISCAN RK-464 eye-tracker. As set out in Figure 6a), clips include complex video stimuli that can be divided into seven categories: *Outdoor, Videogames, Commercials, TV News, Sports, Talk Shows* and *Others*. Eye fixations of at least 4 subjects are provided for each clip. The dataset was delivered some years ago with the same intention pursued with our analysis, and has been employed to evaluate a lot of *state-of-the-art* saliency models. However, to our knowledge, none of them had attempted so far to offer a data interpretation such as the one resulted from our approach.

Hence, in order to both assess the performance and gain insight into the latent information provided by the proposed probabilistic method for visual attention estimation, we will compare two different approaches: a) a *context-generic* (C-G) model trained using frames belonging to videos in all the categories; and b) 7 *context-aware* (C-A) models trained on those videos belonging to each category or genre.

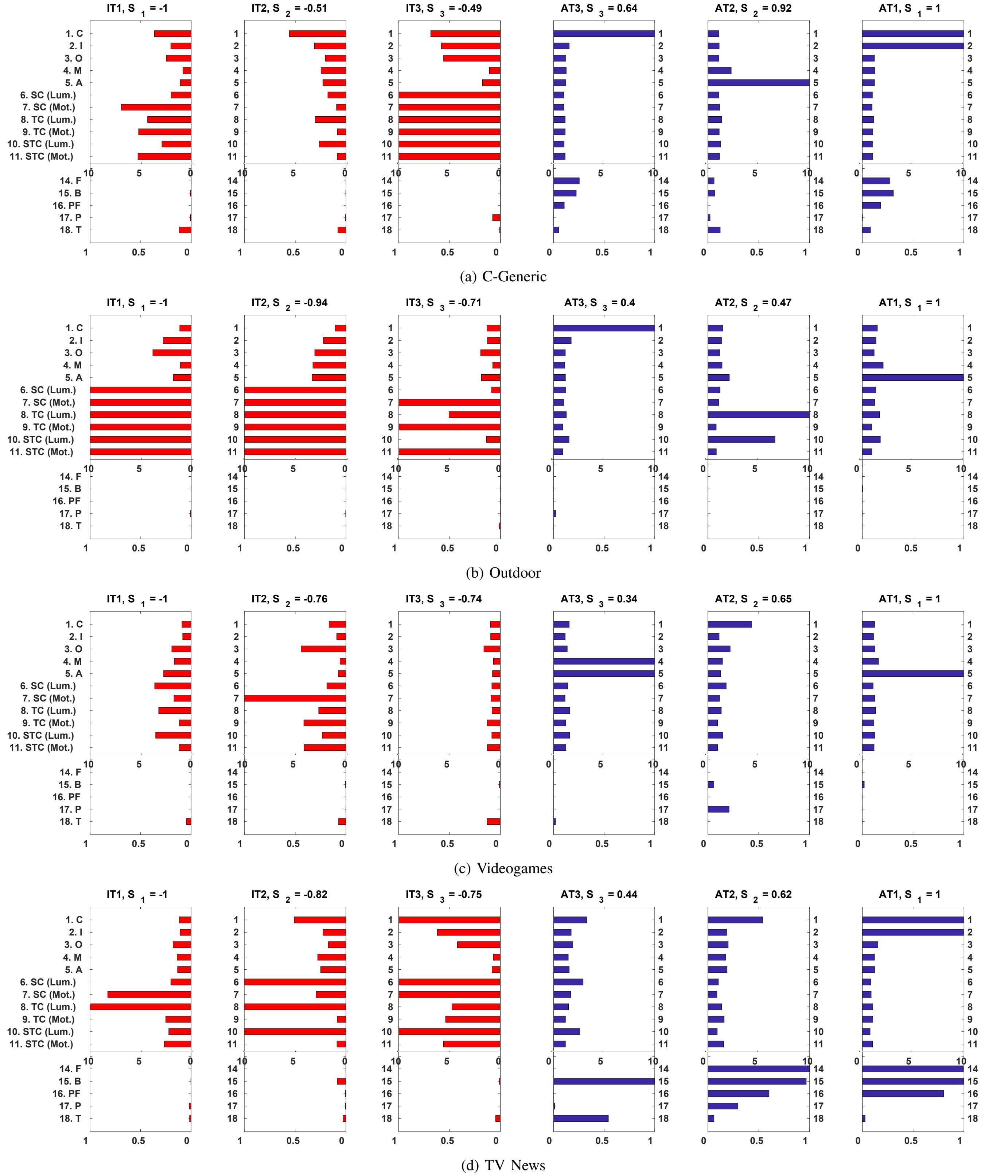
Evaluation of performance

The performance over every video in the dataset is evaluated by conducting a 4-fold cross validation procedure, so that at each iteration some videos are picked for evaluation. For the purpose of avoiding over-fitting, all frames of a video are always grouped together in the same set (train or test).

Bylinskii et al. present in [47] a comprehensive study about visual attention models evaluation, where recommendations for metric selections under specific assumptions and for specific applications are made. According to this extensive analysis, we have selected two suitable metrics to present our results on ground truth fixations prediction. First, for the sake of historical reasons, a ROC-based score is included, which is the *Shuffled Area Under Curve (sAUC)* obtained from the implementation detailed in [48]. Given a video frame, this measure gives less credit to common fixation positions for correct prediction by choosing fixations from different videos as false positives (FP), in such a way that center-bias effects of the spatial distribution of eye fixations are eliminated. However, attention maps that place different amounts of density at fixated locations receive similar scores. The same shuffling technique is applied to the *Shuffled Normalized Scanpath Saliency (sNSS)* metric in [49], which is the second measure we have chosen. A high number of FP drives the overall sNSS down, which makes this metric an interesting supplement to the classical AUC scores. In order to evaluate the performance of visual attention models in a particular video, a probabilistic map that consists of fixations in frames from all other videos in the dataset is used as shuffle map for both scores. 95% confidence bounds are provided for the metrics used.

Model initialization

Due to the stochastic nature of our approach, a correct initialization of the parameters is important to both fasten the convergence and reach an optimal model. As the goal is to learn sub-tasks that either attract or inhibit attention, we initialize basic, novelty and CNNs-based feature distributions as follows: we initialize some topics that inhibit and other that attract visual attention, with $\mu_{kl} = 0$ and $\mu_{kl} = 1$, respectively



(a) C-Generic

(b) Outdoor

(c) Videogames

(d) TV News

Fig. 5. Three most prominent attraction (AT) and inhibition (IT) sub-tasks inferred by (a) *context-generic* and (b) *Outdoor*, (c) *Videogames* and (d) *TV News context-aware* models learned based on CRCNS-ORIG [36] database. Sub-tasks are represented in the graphs as combinations of some of the features described in Section III-B: basic and novelty features, such as *color* (C), *intensity contrast* (I), *orientation* (O), *velocity* (M), *acceleration* (A), *luminance spatial coherence (SC (Lum.))*, *motion spatial coherence (SC (Mot.))*, *lumincance temporal coherence (TC (Lum.))*, *motion temporal coherence (TC (Mot.))*, *luminance spatio-temporal coherence (STC (Lum.))*, *motion spatio-temporal coherence (STC (Mot.))*; and object-based features, such as *frontal* (F) and *profile faces* (PF), *upper bodies* (B), *pedestrians* (P) and *text* (T).

(remember that our features are maps in the range $[0,1]$). Then, in order to provide initial variances for the topics, we compute two separate sets of variances with respect to $\mu_{kl} = \{0,1\}$, from non-attracting and attracting locations respectively. Then, we run a separate *k-means* over the variance values and obtain the corresponding K centroids, one per topic. For camera motion features, the parameters $\mathbf{c}_k$ are randomly initialized with values close to 0 whereas, as we have already mentioned, Σ_k is empirically set to $\Sigma_k = diag(0.25)$. Finally, discrete distribution features for object detection are initialized uniformly for every region in the non-uniform grid.

Last but not least, the main parameter of the proposed model is the number K of sub-tasks or topics that contribute to model visual attention. For simplicity, we have used the same number of attracting and inhibiting topics in our initialization. As indicated in the next sub-section, $K = 60$ is the number of topics used for the rest of the experiments. Finally, initial global topic proportions α have been empirically set to $\alpha_k = 0.01$.

B. Visual attention as a mixture of sub-tasks for generic and context-aware models

The most outstanding outcome of our probabilistic approach is determined by the topics inferred, which effectively help to interpret how visual attention works. Firstly, by means of the proportions in which those are blended, we can establish which sub-tasks are more prevailing for guidance. We have statistically estimated the importance of each topic by examining the value η_k of the logistic regression model and the topic proportions ϕ_{nk} obtained for each spatial location n evaluated on the test set, as both variables are linearly related to the model response which generates the visual attention map. In particular, the relevance score of each sub-task k is computed as:

$$\mathcal{S}_k = \eta_k \sum_{n=1}^{N} \phi_{nk} \tag{16}$$

Scores are later normalized between $[-1, 1]$ to simplify the analysis. Secondly, regarding the distribution parameters learned for features considered as input, we can further study the meaning of sub-tasks, providing useful information about the most conspicuous regions in a given scenario. For the sake of interpretability, it should be noted that we have not considered CNNs-based features in this analysis. Besides, Gaussians' means are not learned and remain fixed in $\mu_{kl} = 0$ and $\mu_{kl} = 1$ during the whole inference process for those topics inhibiting (IT) or attracting attention (AT), respectively. Furthermore, the camera motion distribution has been also removed from the analysis as it has been observed that there is not a strong influence of this feature in any of the categories, since parameters $\mathbf{c}_k$ learned for the most prevailing topics have all similar values. Under this simplified scenario, we can evaluate the relevance of *basic and novelty features*, using their learned standard deviation values σ_{kl} :

$$\mathcal{S}^C_{kl} = \frac{\sigma^F_l}{\sigma_{kl}} \tag{17}$$

with values in the range $[0, +\infty)$. Given a sub-task k, a feature l will be representative if its standard deviation σ_{kl} is lower compared to the deviation σ^F_l measured on areas that correspond with the topic type F (fixated areas if the topic is attracting attention, and viceversa).

Moreover, scores for *object-based features* are calculated by computing the cumulative probability of the cells that lie inside the detected bounding box ($r > 0$, excluding the background cell):

$$\mathcal{S}^D_{kl} = \sum_{r=1}^{R} \beta_{klr} \tag{18}$$

with values between $[0, 1]$.

Scores obtained by the three most noteworthy attraction and inhibition sub-tasks for three video genres are shown in Figure 5. Moreover, significant sub-tasks deduced by the *context-generic* model are provided for the sake of comparison. Although the number of topics experimentally determined is quite high ($K = 60$), we have observed that only few of them are responsible of guiding attention most of the time, whereas the rest are intended to refine the estimation, specially in the less prevalent sequences.

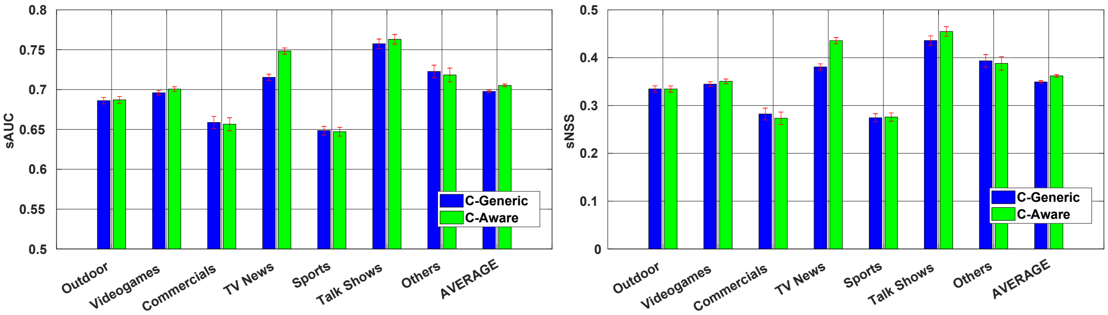
As can be seen, different sub-tasks are determined to model visual attention in each scenario, existing an appreciable contrast between well-separated categories such as *Outdoor* or *TV News*, which involve distinctive actions. While *context-generic* model is adjusted to the most prominent events in the whole database, which consist of faces noticeable by their color and intensity, and motion objects, *context-aware* models have the ability of attaining more particular and explainable activities. Motion and acceleration features are relevant in *Outdoor* and *Videogames* sub-tasks, which could be related to people or characters walking or running. In contrast, faces and texts are more attractive and predominant in categories like *Commercials*, *TV News* and *Talk Shows*. Both motion and faces are eye-catching in *Sports* videos, which could be understood as a combination of these two first mentioned types. Last but not least, low values of spatial and temporal coherency features are mostly frequent in IT, which implies reducing the attentional response in usual and stable locations over space and time.

C. Results on the CRCNS-ORIG database

In this second set of experiments, CNNs-based features are included and the ATOM model learns unconstrained Normal distributions without fixating the means. Results obtained for the two versions of our method in each category are provided in Figure 6b). As can be seen, the *context-aware* models match or outperform the *generic* approach in all genres. Without considering *Others* category, which is more diverse and contains a synthetic saccade test video, best scores are obtained for *TV News* and *Talk Shows* genres, due to the proper operation of object detectors incorporated as input for the model, as shown in some of the examples provided in Figure 7. Scores achieved for *Outdoor* and *Videogames* videos are also remarkable, due to the strong influence assigned to motion-related features. This reinforces the idea that, depending on the context, certain particular sub-tasks aid to guide visual attention. This can be also noticed if we look at the results obtained in a category whose associated videos are not closely related, such as *Others*, where it has been hard to find out meaningful

Context	# clips	Frames
Outdoor	17	8,357
Videogames	9	15,809
Commercials	4	2,618
TV News	7	8,071
Sports	5	4,851
Talk Shows	4	4,244
Others	4	2,539
TOTAL	50	46,489

(a)



(b)

Fig. 6. (a) Categories into which the CRCNS-ORIG [36] database is divided. (b) Results obtained by the proposed *context-generic* and *context-aware* ATOM models, which consist of $K = 60$ topics.

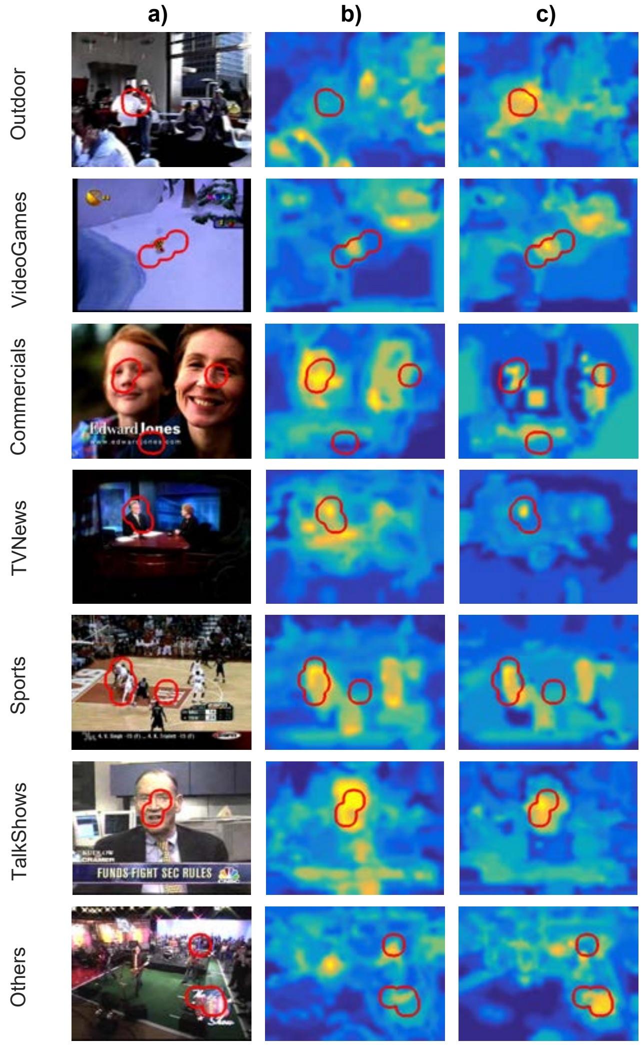


Fig. 7. Visual attention maps obtained by ATOM for some example frames from CRCNS-ORIG [36] database. Red boundaries highlight high-density regions of human fixations in the GT map. (a) Original frames. (b) Context-Generic. (c) Context-Aware.

topics. In fact, the results in this case for the *context-generic* model are higher, probably because it has been trained on a wider set of videos similarly related than those used to obtain its corresponding *context-aware* approach, which might have allowed for a better generalization. Therefore, it can be concluded that it is necessary to establish well-defined application scenarios where to determine these feature-based representations. In order to provide a fair comparison, we draw on the same number of topics for each of the categories in the dataset chosen, although it has been observed that the performance depends on the complexity of the scenarios. If we compare the average performance of *context-aware* models with respect to the result obtained by the *context-generic* approach, there is an improvement of 4.1% in terms of sNSS and 1.2% in terms of sAUC, which is closer to the upper threshold given by H50 score. Thus, we can state that specific *context-aware* representations of visual attention learned over smaller training sets (the training videos belonging to each category) work better than *generic* models over larger datasets (including all video categories). Based on these results, from now on we will use the *context-aware* version of our algorithm to provide a comparison with other approaches in the state-of-the-art.

D. *Comparison with state-of-the-art methods*

With the aim of assessing the performance of our approach in comparison with other methods available in the state-of-the-art, we have selected seventeen static and dynamic visual attention models, which are representative of the existing diversity for visual attention prediction: we have included both BU and TD or learnable models, a model that uses CNNs to predict, etc., as well as three reference models introduced in [50] (H50, CHANCE, CENTER). Parameters used are the ones set as default by authors. As can be verified from CENTER and CHANCE baselines, all metrics included in the analysis are not affected by center bias effect.

Table I contains all the results obtained for the assessed methods, together with those reached by the system introduced in this article (ATOM). We also include on the list the first approach we presented in [13], which make use of a linear regressor to estimate visual attention instead of the logistic regressor currently employed. Features and number of topics ($K = 40$) taken for this previous configuration are those reported in [13].

The improvement achieved by our model with respect to very recent approaches such as AWS-D [49], DCL [32], WMAP [51] or ICL-D [52] is statistically significant. Moreover, it is also visually noticeable in some intricate cases, as those shown in Figure 8, with scenes showing crowds, multiple similar concepts that hamper visual guidance or quick actions.

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS IN THE CRCNS-ORIG [36] DATABASE.

Model	Learning	$sAUC$ $mean\ (C.I.)^{Rank}$	$sNSS$ $mean\ (C.I.)^{Rank}$
ATOM	YES	$\mathbf{0.705\ (0.703, 0.707)^{1}}$	$\mathbf{0.362\ (0.359, 0.365)^{1}}$
AWS-D [49]	NO	$0.700\ (0.698, 0.702)^{2}$	$0.322\ (0.319, 0.325)^{3}$
DCL [32]	YES	$0.684\ (0.682, 0.686)^{3}$	$0.323\ (0.320, 0.326)^{2}$
AWS [53]	NO	$0.675\ (0.674, 0.677)^{4}$	$0.281\ (0.278, 0.285)^{4}$
WMAP [51]	NO	$0.670\ (0.669, 0.672)^{5}$	$0.236\ (0.232, 0.239)^{12}$
Hou and Zhang [54]	NO	$0.669\ (0.667, 0.671)^{6}$	$0.260\ (0.257, 0.263)^{7}$
DCL+ [32]	YES	$0.666\ (0.665, 0.668)^{7}$	$0.255\ (0.251, 0.258)^{8}$
ICL-D [52]	NO	$0.666\ (0.665, 0.668)^{8}$	$0.217\ (0.214, 0.220)^{14}$
PQFT [55]	NO	$0.662\ (0.660, 0.663)^{9}$	$0.243\ (0.240, 0.246)^{11}$
Goferman [56]	NO	$0.661\ (0.659, 0.662)^{10}$	$0.263\ (0.260, 0.266)^{6}$
SUN [27]	YES	$0.654\ (0.652, 0.655)^{11}$	$0.251\ (0.248, 0.254)^{9}$
AIM [2]	YES	$0.653\ (0.652, 0.655)^{12}$	$0.270\ (0.268, 0.273)^{5}$
Torralba [57]	NO	$0.648\ (0.646, 0.650)^{13}$	$0.251\ (0.248, 0.254)^{10}$
Itti (ST) [19] [20]	NO	$0.634\ (0.632, 0.636)^{14}$	$0.217\ (0.214, 0.220)^{15}$
Fernández-Torres [13]	YES	$0.628\ (0.626, 0.630)^{15}$	$0.218\ (0.215, 0.221)^{13}$
SDSR [58]	NO	$0.627\ (0.625, 0.628)^{16}$	$0.129\ (0.126, 0.132)^{17}$
GBVS (ST) [20]	NO	$0.621\ (0.619, 0.623)^{17}$	$0.182\ (0.179, 0.186)^{16}$
ESA-D [59]	*NO*	$0.541\ (0.539, 0.543)^{18}$	$0.075\ (0.072, 0.078)^{18}$
H50	*NO*	$0.800\ (0.799, 0.802)$	$0.679\ (0.677, 0.681)$
CHANCE	*NO*	$0.500\ (0.500, 0.500)$	$-0.000\ (-0.000, 0.000)$
CENTER	*NO*	$0.509\ (0.507, 0.511)$	$0.057\ (0.054, 0.060)$

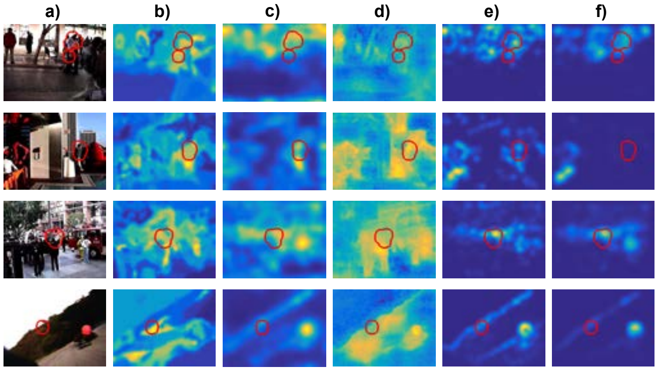


Fig. 8. Visual attention maps generated by some of the most outstanding methods in the *state-of-the-art* for some intricate example frames taken from CRCNS-ORIG [36] database. Red boundaries highlight high-density regions of human fixations in the GT map. (a) Original frames. (b) ATOM. (c) AWS-D [49]. (d) DCL [32]. (e) WMAP [51]. (f) ICL-D [52].

Finally, we evaluate the computational time of the test phase for all methods on a system with an Intel Core i7-6700K CPU with 4.00GHz. Regarding our approach, we should distinguish between the learning and the test phase. Both phases involve a feature extraction stage that takes $5.81s$ per frame. Time spent in the learning phase depends on the number of topics of the model trained and the amount of input frames. For instance, training a model with $K = 60$ topics and ~ 3000 frames would take $\sim 45min$. This time can be reduced if the number of topics is decreased to $K = 40$ ($\sim 32min$) or $K = 20$ ($\sim 18min$), which would slightly decrease the performance. Then, in the test phase, the average time per frame is only $0.157s$, which is competitive compared to those obtained by the two next best methods, AWS-D [49] ($0.075s$) and DCL [32] ($0.2s$).

E. Where we are: model strengths and limitations

Despite the improvement reached by the proposed model over the *state-of-the-art* and the compelling information it provides, we are still far from reaching human capacity of almost immediately selecting the most essential elements and areas to reach a full understanding in a given scenario, or to solve a particular task. H50 score reflected in Table I, which is calculated for each frame by means of a map that contains the fixations of the 50% of subjects available, constitutes a good realistic upper threshold to put into perspective the efficiency of the existing approaches. Nonetheless, we advocate that the inclusion of an intermediate level between features and visual attention in terms of sub-tasks is a powerful way towards comprehensible guiding representations. In order to assess the influence of the topic models over the final result, we have evaluated an alternative method that uses a logistic regressor over the same set of features to directly predict visual attention. Our topic model achieves a relative improvement of 22.3% in terms of sNSS and 1.7% in terms of sAUC. This clearly demonstrates that the topic-based hierarchical modeling is useful, not only because it provides meaningful representations of top-down visual attention, but also because it successfully enhances the system performance.

We have demonstrated that some of the traditional basic features used (e.g. color, orientation, motion) continue being useful in many cases to predict visual attention in videos. Furthermore, thanks to the object detectors introduced and the corresponding spatial discrete distributions, we are able to model simple but attractive concepts such as faces or text, putting emphasis on their most noticeable elements. The high performance achieved by these detectors in some categories leads us to reckon the integration of large-scale hierarchical networks for object recognition in future revisions of our model, such as the ones evaluated in the ImageNet Challenge [60]. On the other hand, there is also a need of a deeper understanding of the scene, establishing relations between recognized concepts both in the same frame or in different frames. This would enable the system to enhance guidance in situations where many conspicuous regions exist and it is required to select the most significant, or even an intermediate one (e.g. Figure 9a)); when objects are occluded during few frames (e.g. Figure 9b)); or to determine the sequence of objects or subjects to follow in order to interpret a scene (e.g. Figure 9c)), among others. In other words, we pursue the identification and modeling of sub-tasks, not only over space but also along time.

Last but not least, it has to be mentioned the importance of GT eye fixations, both in learning and evaluation stages. As can be appreciated in some of the examples gathered throughout the article, not all fixations contain useful information to train a visual attention system, not only because the occlusions mentioned above, but also due to errors derived from eye-tracker capture or the observers' center bias present in numerous frames. Fixations often fall on edges, not covering completely some objects of interest, such as gaming characters or players, which are essential to infer sub-tasks. Additionally, we might take into account covert

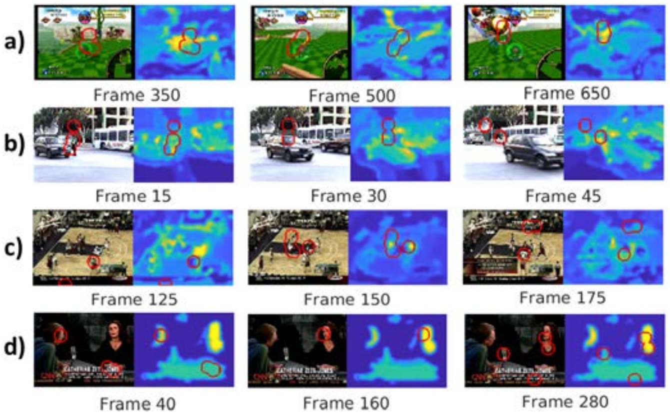


Fig. 9. Frame sequences taken from CRCNS-ORIG [36] database to analyze some ATOM model drawbacks and define future lines of research. Red boundaries highlight high-density regions of human fixations in the GT map, both in original frames and computed visual attention maps. (a) Videogames scenario where many remarkable regions exist, making observers constantly shift their gaze. (b) Outdoor scenario where multiple salient concepts (e.g. car, policeman) overlap each other. (c) Basketball match, in which the sequence of players to follow is decisive to model visual attention. (d) TV talk show, where several quasi-static concepts appear together during a long time lapse and estimated visual attention is either distributed among all or focused in one of them.

attention, which is independent of eye movements and stresses the existence of attention independent of gaze change. Hence, techniques to filter and, if necessary, extend regions considered as GT should be regarded in upcoming experiments. What is more, existing evaluation metrics do not seem to be appropiate in situations such as the one shown in Figure 9d), where many remarkable quasi-static concepts appear together during a long time lapse and estimated visual attention is either distributed among all or focused in one of them. If observers' fixations are widely dispersed and continuously displaced between their corresponding locations, what should be the GT taken for each frame in this case? Should all concepts be considered as attractive during the whole video fragment? We will seek to address these issues in future application scenarios.

V. Conclusions

In this paper, we have presented a hierarchical probabilistic framework to estimate and understand TD visual attention in videos. Relying on the idea of 'guiding representation' supported by some of the most prevailing psychological theories about visual attention, our ATOM model decomposes it into mixtures of several latent topics or sub-tasks, which are in turn modeled as combinations of low-, mid- and high-level spatio-temporal features obtained from video frames. For that purpose, an intermediate level between feature extraction and visual attention computation phases is introduced, aligning the latent discovered sub-tasks from frames to the information drawn from human fixations. The attention response is thus generated by computing a logistic regression model over topic proportions. It is also worth mentioning that the definition of the method is generic and independent of the input features, which enables an easy adaptation to any application scenario.

The ability of ATOM of successfully learning specifically adapted hierarchical representations of visual attention in diverse contexts has been demonstrated on the basis of a wide set of features. Either classical and easily interpretable feature maps, which have been effective to extract conclusions about the existing scenarios in the well-known CRCNS-ORIG [36] database, or those generated by recently proposed CNNs structures, which allow to capture more complex concepts, have aided to significantly outperform other competent methods in the literature. Moreover, the detection of simple elements such as faces or text, and their modeling as spatial discrete distributions, has led to improve visual attention estimation in certain challenging situations.

Experimental results show the advantage of obtaining comprehensible guiding representations to model visual attention. However, it is still necessary to deepen in some of the stages of the framework, carefully selecting the most meaningful information from fixated regions in the scene, and integrating more robust recognition and understanding techniques that enable to identify more accurate sub-tasks over space and time. To that end, future efforts will be directed towards task-driven approaches, developing video databases with human fixations to test the usefulness of the system in end-user applications.

Acknowledgment

This work has been partially funded by the Spanish Ministry of Education, Culture and Sports FPU Grant FPU13/04904, and the National Grants TEC2014-53390-P and TEC2014-61729-EXP of the Spanish Ministry of Economy and Competitiveness.

References

[1] J. K. Tsotsos, *A Computational Perspective on Visual Attention*, 1st ed. The MIT Press, 2011.

[2] N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *NIPS*, 2005.

[3] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005)*. Cambridge, MA: MIT Press, 2006, su;mod;bu;td;eye, pp. 547–554.

[4] N. Sprague and D. Ballard, "Eye movements for reward maximization," in *In Advances in Neural Information Processing Systems 15*. MIT Press, 2003.

[5] Z. Ren, S. Gao, L. T. Chia, and I. W. H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 769–779, May 2014.

[6] T. V. Nguyen, Z. Song, and S. Yan, "Stap: Spatial-temporal attention-aware pooling for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 1, pp. 77–86, Jan 2015.

[7] T. Yubing, F. A. Cheikh, F. F. E. Guraya, H. Konik, and A. Trémeau, "A spatiotemporal saliency model for video surveillance," *Cognitive Computation*, vol. 3, no. 1, pp. 241–263, 2011.

[8] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proceedings of the Tenth ACM International Conference on Multimedia*, ser. MULTIMEDIA '02. New York, NY, USA: ACM, 2002, pp. 533–542.

[9] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982, July 2011.

[10] A. M. Treisman and G. Gelade, "A feature-integration theory of attention." *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[11] J. M. Wolfe, "Guided search 2.0 a revised model of visual search," *Psychonomic bulletin & review*, vol. 1, no. 2, pp. 202–238, 1994.

[12] A. L. Yarbus, *Eye Movements and Vision*. Plenum. New York., 1967.

[13] M. A. Fernández-Torres, I. González-Díaz, and F. D. de María, "A probabilistic topic approach for context-aware visual attention modeling," in *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, June 2016, pp. 1–6.

[14] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Reviews Neuroscience*, vol. 5, no. 6, pp. 495–501, 2004.

[15] J. M. Wolfe, "Guided search 4.0," *Integrated models of cognitive systems*, pp. 99–119, 2007.

[16] J. M. Henderson and A. Hollingworth, "High-level scene perception," *Annual review of psychology*, vol. 50, no. 1, pp. 243–271, 1999.

[17] A. Borji and L. Itti, "Defending yarbus: Eye movements reveal observers' task," *Journal of vision*, vol. 14, no. 3, pp. 29–29, 2014.

[18] 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.

[19] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[20] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19*. MIT Press, 2007, pp. 545–552.

[21] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[22] N. Sprague and D. Ballard, "Eye movements for reward maximization," in *Advances in neural information processing systems*, 2003, p. None.

[23] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2006, pp. 2049–2056.

[24] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[25] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 2106–2113.

[26] L. Elazary and L. Itti, "A bayesian model for efficient visual search and recognition," *Vision Research*, vol. 50, no. 14, pp. 1338 – 1352, 2010, visual Search and Selective Attention.

[27] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, p. 32, 2008.

[28] J. Li, Y. Tian, T. Huang, and W. Gao, "Multi-task rank learning for visual saliency estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 5, pp. 623–636, May 2011.

[29] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2798–2805.

[30] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1072–1080.

[31] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?" in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, 2016, pp. 809–824.

[32] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 478–487.

[33] Ç. Bak, A. Erdem, and E. Erdem, "Two-stream convolutional networks for dynamic saliency prediction," *CoRR*, vol. abs/1607.04730, 2016.

[34] A. Palazzi, F. Solera, S. Calderara, S. Alletto, and R. Cucchiara, "Where should you attend while driving?" *CoRR*, vol. abs/1611.08215, 2016.

[35] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level cnn: Saliency-aware 3-d cnn with lstm for video action recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 510–514, April 2017.

[36] L. Itti and R. Carmi, "Eye-tracking data from human volunteers watching complex video stimuli," Dec 2009.

[37] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[38] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 121–128.

[39] S.-H. Yang, H. Zha, and B.-G. Hu, "Dirichlet-bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora," in *Advances in neural information processing systems*, 2009, pp. 2143–2150.

[40] Z. Kato and T.-C. Pong, "A markov random field image segmentation model for color textured images," *Image and Vision Computing*, vol. 24, no. 10, pp. 1103 – 1114, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885606001223

[41] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 5 2009.

[42] D. Mahapatra, S. O. Gilani, and M. K. Saini, "Coherency based spatio-temporal saliency detection for video object segmentation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 3, pp. 454–462, June 2014.

[43] G. Abdollahian, Z. Pizlo, and E. J. Delp, "A study on the effect of camera motion on human visual attention," in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 693–696.

[44] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–511.

[45] C. Harris and M. Stephens, "A combined corner and edge detector." Citeseer, 1988.

[46] I. González-Díaz, V. Buso, and J. Benois-Pineau, "Perceptual modeling in the problem of active object recognition in visual scenes," *Pattern Recognition*, vol. 56, pp. 129–141, 2016.

[47] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *arXiv preprint arXiv:1604.03605*, 2016.

[48] J. Zhang and S. Sclaroff, "Exploiting surroundedness for saliency detection: a Boolean map approach," *IEEE Trans. Pattern Anlaysis and Machine Intellegence (TPAMI)*, 2015.

[49] V. Leboran, A. Garcia-Diaz, X. Fdez-Vidal, and X. Pardo, "Dynamic whitening saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2016.

[50] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," in *MIT Technical Report*, 2012.

[51] F. López-García and X. M. Pardo, *Scene recognition through visual attention and image features: A comparison between sift and surf approaches*.

[52] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in neural information processing systems*, 2009, pp. 681–688.

[53] A. Garcia-Diaz, V. Leborn, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *Journal of Vision*, vol. 12, no. 6, p. 17, 2012.

[54] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[55] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, Jan 2010.

[56] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, Oct 2012.

[57] A. Torralba, "Modeling global scene factors in attention," *J. Opt. Soc. Am. A*, vol. 20, no. 7, pp. 1407–1418, Jul 2003.

[58] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, p. 15, 2009.

[59] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proceedings of the 11th European Conference on Computer Vision: Part V*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 366–379.

[60] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

Miguel-Ángel Fernández-Torres received the Audiovisual Systems Engineering degree from Universidad Carlos III de Madrid, Madrid, Spain, in 2013, and the Master degree in Multimedia and Communications from Universidad Carlos III de Madrid, Spain, 2014. He is currently pursuing his Ph.D. degree at the Signal Theory and Communications Department in Universidad Carlos III de Madrid, Madrid, Spain. His current research interests include visual attention modeling, image and video analysis, medical image classification, and computer vision.

Iván González-Díaz received the Telecommunications Engineering degree from Universidad de Valladolid, Valladolid, Spain, in 1999, the M.Sc. and Ph.D. degree from Universidad Carlos III de Madrid, Madrid, Spain, in 2007 and 2011, respectively. After holding a postdoc position in the Laboratoire Bordelais de Recherche en Informatique at the University Bordeaux, he currently works as a Visiting Lecturer at the Signal Theory and Communications Department in Universidad Carlos III de Madrid. His primary research interests include object recognition, category-based image segmentation, scene understanding and content-based image and video retrieval systems. In these fields, he is co-author of several papers in prestigious international journals, two chapters in international books and a few papers in revised international conferences.

Fernando Díaz-de-María received the Telecommunication Engineering degree and the Ph.D. degree from the Universidad Politécnica de Madrid, Madrid, Spain, in 1991 and 1996, respectively. Since October 1996, he has been an Associate Professor in the Department of Signal Processing and Communications, Universidad Carlos III de Madrid, Madrid, Spain. His primary research interests include video coding, image and video analysis, and computer vision. He has led numerous projects and contracts in the fields mentioned. He is co-author of numerous papers in peer-reviewed international journals, several book chapters and a number of papers in national and international conferences.