A Probabilistic Topic Approach for Context-Aware Visual Attention Modeling

Miguel-Ángel Fernández-Torres, Iván González-Díaz, Fernando Díaz-de-María
Department of Signal Theory and Communications
Universidad Carlos III de Madrid
Avda. de la Universidad 30, Leganés, Madrid, Spain
{matorres, igonzalez, fdiaz}@tsc.uc3m.es

***Abstract*—The modeling of visual attention has gained much interest during the last few years since it allows to efficiently drive complex visual processes to particular areas of images or video frames. Although the literature concerning bottom-up saliency models is vast, we still lack of generic approaches modeling top-down task and context-driven visual attention. Indeed, many top-down models simply modulate the weights associated to low-level descriptors to learn more accurate representations of visual attention than those ones of the generic fusion schemes in bottom-up techniques. In this paper we propose a hierarchical generic probabilistic framework that decomposes the complex process of context-driven visual attention into a mixture of latent subtasks, each of them being in turn modeled as a combination of specific distributions of low-level descriptors. The inclusion of this intermediate level bridges the gap between low-level features and visual attention and enables more comprehensive representations of the later. Our experiments on a dataset in which videos are organized by genre demonstrate that, by learning specific distributions for each video category, we can notably enhance the system performance.**

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

I. Introduction

The interest of understanding how humans perceive and react to the great number of visual stimuli in the real world has motivated to many researchers during the last years. A region of a scene can draw our attention at the same time as others are ignored. Observers are generally attracted by the most informative [1] and surprising [2] regions, and also by those that, given a task, allow to achieve an objective [3].

The huge amount of visual data available nowadays constitutes an important source of information that requires of automatic analysis techniques in order to select spatial and temporal regions of special interest. Visual attention models have been proposed for that purpose in many different image and video applications, such as object [4] and action [5] recognition, video surveillance [6], image retrieval [7], or video summarization [8].

Two main families of visual attention models can be distinguished: bottom-up and top-down. On the one hand, bottom-up (BU) models are mainly based on characteristics of the visual scene (*stimulus-driven*) such as color, orientation, movement or depth. BU attention is fast, involuntary and mostly feedforward. On the other, top-down (TD) models (*goal-driven*) are determined by cognitive phenomena like knowledge, expectations or goals. In these models, attention is focused on advanced indications. TD attention, in contrast to BU, is slower, task-driven, voluntary and closed-loop.

The origin of many of the existing visual attention models is the Treisman and Gelade's "Feature Integration Theory" [9], which introduced relevant features for the perception of objects. Koch and Ullman [10] designed a model to combine these 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.

Borji et al. presented a survey on visual attention modeling in [11]. The first implementation and verification of a BU model, which uses color, intensity and orientation features, was performed by Itti et al. [12]. Harel et al. [13] proposed a saliency algorithm based on graphs, which extracted the same features at different scales. These two representations are the most frequently employed due to their good performance in a lot of situations.

While a large amount of BU models have been developed, TD methods are still scarce, and generally integrated within systems particularly tailored to very specific scenarios. In these cases, only the efficiency of the whole scheme is evaluated. Moreover, although the prevailing view is that both BU and TD mechanisms cooperate to guide our attention, few investigations address this cooperative approach.

To overcome these limitations, we present a hierarchical probabilistic framework to estimate visual attention in videos, which can be applied to different scenarios. In our model TD visual attention is decomposed into mixtures of various latent sub-tasks that are in turn modeled as combinations of low-level features. According to each scenario, distinct features could draw visual attention: e.g. in sports videos, for instance, motion features will be needed to follow players, while color or intensity can highlight some objects in video games scenes or ads. We will show how our approach successfully learns particularly adapted hierarchical representations of visual attention in various scenarios (video genres in our case), thus improving the performance of a generic system trained using all data.

The remainder of this paper is organized as follows: Section II reviews the related work, focusing mainly on TD approaches. Section III describes the generative model proposed, the features extracted and a novel normalization procedure.

Experimental results are gathered in Section IV, and Section V summarizes our conclusions and outlines further work.

II. Related Work

As mentioned above, despite their importance in the process of driving visual attention, we still lack of generic TD approaches. In general, most TD models guide attention towards specific targets by modulating gains associated with low-level stimuli. Navalpakkam and Itti [14] optimized the integration of BU cues for target detection by maximizing the signal-to-noise ratio of the target vs. background. Elazary and Itti [15] proposed a more flexible model that can both select the best features to guide attention and adjust the width of feature detectors. The evaluation of TD algorithms is often performed at application level, such as in [16], [17].

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 [18] a probabilistic model that defines saliency as the pointwise mutual information between BU local features and TD search target features. Li et al. [19] 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.

In contrast to [19], our hierarchical model introduces a new intermediate level between feature extraction and visual attention computation stages. This level is defined by means of *latent topics* that represent a set of subtasks that contribute to the complex process of visual attention. Latent topics are not directly observed but unsupervisedly inferred as combinations of low-level features. In addition, the top level of our approach is related to visual attention, which is modeled as a linear regression over the topic proportions at each spatial location. The weights associated with the regressor will be trained using human fixations as training data.

The definition of the model is generic and independent of the application scenario and can therefore be seamless adapted to any scenario of application by learning from the expert/human fixations. Furthermore, 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.

III. A Generative Model for Visual Attention

A. Overview

The proposed model relies on the following assumption: *Task-driven visual attention in video can be modeled as a mixture of several subtasks which, in turn, can be represented as combinations of low-level spatio-temporal features obtained from video frames*. For example, in a video-surveillance scenario, the visual attention of an operator may be attracted by various events: objects/people moving extremely fast or with irregular trajectories, abandoned objects, explosions, crowds, accesses to restricted areas, etc. Let us note that we do not aim to detect the final events of interest for an application, but to efficiently guide the later processing to particular areas of special importance in the video. Hence, the goal of our proposal is to automatically identify a set of subtasks that contribute to the process of visual attention in a particular context or application and, moreover, to model these subtasks as combinations of low-level spatio-temporal features. For that end, we propose a probabilistic Latent Topic Model (LTM) approach based on the well-known *Latent Dirichlet Allocation* (LDA) [20] method and its supervised extension [21].

In our particular context, task-driven visual attention is modeled as a finite mixture over a set of K topics $\mathbf{z} = \{z_1, z_2, ..., z_K\}$, which represent the subtasks contributing to attention. Then, for a given video frame I_t, a set of L low-level visual descriptors $\mathbf{f} = \{f_1, f_2, ..., f_L\}$ is extracted at each spatial location. The topics are in turn modeled as combinations of these low-level features. For simplicity, we will consider conditional independence among features, so that the joint distribution of features for a particular topic can be factorized into the individual probability distributions $p(f_l|z, \Gamma_l)$, where Γ_l stands for parameters of the distributions.

The original LDA infers topics in an unsupervised way, discovering an underlying structure from a collection of documents. However, as our system aims to learn how humans guide their attention to some visual stimuli, we will drive our training step using ground-truth (GT) fixations provided by different subjects. Since the proposed system aims to learn how humans guide their attention to some visual stimuli, it relies on the supervised version of the LDA [21]. The *Supervised Latent Dirichlet Allocation* (sLDA) associates a continuous response variable y with each document, and then learns a linear model over the latent topics proportions to predict that variable y. In our approach, as the goal is to predict a response variable (the visual attention) for each spatial location $x \in X$ in a frame, we will consider an alternative response variable g_x at each location x.

In the next subsections we describe the set of input visual features extracted and the generative LTM proposed.

B. Feature Extraction and Normalization

The proposed framework naturally incorporates a great number of diverse features. Depending on their nature, it is possible to model them using various probability density functions: e.g. *normal*, *exponential*, *discrete*, etc. In order to provide a sufficiently rich and valid representation suitable for a variety of scenarios, 11 low-level features have been considered.

- **Color**, **Intensity** and **Orientation** maps proposed by Itti et al. in [12].
- **Velocity and Acceleration**: for each spatial location x, we first compute motion vectors using the optical flow method provided in [22]. Then, the motion magnitude M_x (using the L2-norm) and its absolute derivative A_x are computed.
- **Coherency-Based Features**: coherency features relies on the distribution of pixel values over space and time to highlight those regions where dispersion is large and

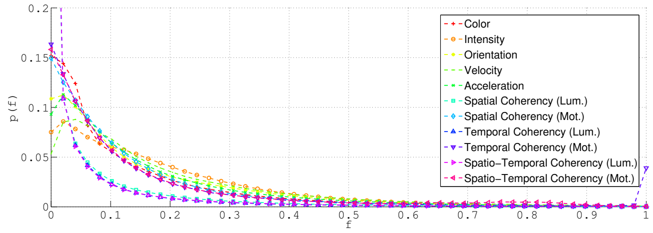

Fig. 1. Sample distributions for the feature maps considered. All distributions $p(f_l)$ can be approximated by Gaussians with zero mean and a given standard deviation σ_l.

might be more salient for observers. We draw on the work done in [23], but use the *quartile coefficient of dispersion* (QCD) as dispersion statistic, which is computed as $QCD = \frac{Q_3 - Q_1}{Q_3 + Q_1}$, being Q_1 and Q_3 the first and third quartiles, respectively. *Spatial, temporal* and *spatio-temporal* coherency maps are extracted. In total, we compute 6 maps: three over the pixel intensity values and three over the motion magnitude.

Feature Normalization

A normalization step is crucial to generate suitable feature maps to be properly fused in our probabilistic framework. Hence, we have designed two types of normalization: a) *statistical feature normalization* and b) *inter-feature intra-frame normalization*.

Figure 1 shows the sample distributions for the considered feature maps. Since all distributions $p(f_l)$ can be approximated by Gaussians with zero mean and a given standard deviation σ_l, we have developed a *Statistical Feature Normalization* (SFN) as follows: first, we approximate each feature distribution by $\mathcal{N}(0, \sigma_l^2)$; second, we normalize it using a multiple of the standard deviation $\hat{f}_l = \frac{f_l}{6\sigma_l}$; finally, we truncate the maximum value to 1 (which corresponds to $6\sigma_l$, and covers the ~ 99.7 of the data samples).

In addition, we have observed that we often find maps where very large areas of the frame show high values of a feature ($f_l \sim 1$). We know that this behavior does not correlate well with the idea of visual attention, which cannot be shared all along the frame. Furthermore, we have also noticed that these features tend to dominate the inference process, thus causing a drop of the performance. Consequently, assuming that visual attention usually focuses on small areas (the fixations), we have proposed the following *inter-feature intra-frame normalization* (IF-IF):

$$\hat{f}_{lx} = \frac{f_{lx}}{C \exp\left(\gamma(\sum_{x=1}^{X} f_{lx} - 1)\right)} \tag{1}$$

where γ has been heuristically set to $\gamma = 1e-4$; C has been computed as $C = max(\hat{f}_{1x}, \hat{f}_{2x} ... \hat{f}_{Lx})$, ensuring that at least one stimulus l is maximal at one spatial location x ($\hat{f}_{lx} = 1$ for one x, l). This assumption seems reasonable if we provide a sufficiently rich and expressive set of features.

C. Proposed Supervised Topic Model

The proposed LTM involves the following generative process for each frame I_t in a corpus of videos $\mathcal{I} = \{I_1, I_2, ..., I_T\}$:

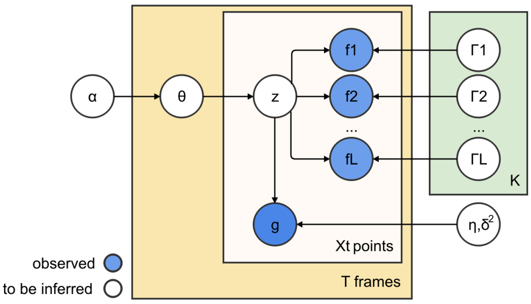

Fig. 2. Graphical representation of the proposed generative topic model for visual attention. Shaded circles represent observations from frames, white circles indicate hidden variables to be inferred, and boxes mean independent repetitions.

1) Draw corpus global proportions for the K topics using a Dirichlet distribution: $\theta | \alpha \sim Dir(\alpha)$.
2) For each spatial location $x \in X$ in the frame:
   a) Draw topic assignment using a multinomial distribution: $z_x | \theta \sim Mult(\theta)$.
   b) Represent the local appearance of the spatial location x by drawing L visual features using Gaussian distributions: $f_{lx} | z_x, \Gamma_{lz_x} \sim N(\mu_{lz_x}, \sigma^2_{lz_x})$.
   c) Draw the response variable modeling the visual attention using a linear regression model: $g_x | z_x, \eta, \delta^2 \sim \mathcal{N}(\eta^T z_x, \delta^2)$.

A graphical representation of the model is shown in Figure 2. Intuitively, the K latent topics represent the subtasks that contribute to visual attention. Let us note that some of these subtasks may attract human attention whereas others may inhibit it. Hence, for each video frame, 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 so, for each x, we first select a subtask by using the index-variable z_x (z_x is K-length vector with a 1 on the location associated with the selected topic). Based on z_x, we draw the local appearance of the spatial location using the particular feature-topic distribution $f_{lx} | z_x \Gamma_{lz_x}$. Finally, we also generate the attention response g_x by computing the linear regression model over the selected topic.

As in the original sLDA [21], exact inference is not possible due to the coupling between the variables θ and 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:

$$q(\theta, z | \gamma, \phi_{1:X}) = q(\theta | \gamma) \prod_{x=1}^{X} q(z_x | \phi_x) \tag{2}$$

that incorporates two new variational parameters: ϕ, which is the parameter of a multinomial distribution $q(z_x | \phi_x)$, 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:

$$
\begin{aligned}
&log\ p(f_{1:X,1:L}, g_{1:X}|\alpha, \Gamma_{1:L,1:K}, \eta, \delta^2) \geq E_q[log\ p(\theta|\alpha)] \\
&+ \sum_{x=1}^{X} E_q[log\ p(z_x|\theta)] + \sum_{x=1}^{X} E_q[log\ p(f_{x,1:L}|z_x, \Gamma_{1:L,1:K})] \\
&+ \sum_{x=1}^{X} E_q[log\ p(g_x|z_x, \eta, \delta^2)] + H(q) \qquad (3)
\end{aligned}
$$

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. (3) and the entropy of the variational distribution are identical to the corresponding terms in the ELBO for unsupervised LDA and are described in [20]. The third term is the expected log probability of the data given the related topic model parameters. Assuming Gaussian distributions $\Gamma_{l,1:K} \sim \{\mu_{l,1:K}, \sigma^2_{l,1:K}\}$ and conditional independence between features:

$$
\begin{aligned}
&E_q[log\ p(f_{x,1:L}|z_x, \Gamma_{1:L,1:K})] = -\sum_{k=1}^{K} \phi_{xk} \sum_{l=1}^{L} \log(\sigma_{kl}\sqrt{2\pi}) \\
&- \sum_{k=1}^{K} \phi_{xk} \sum_{l=1}^{L} \frac{(f_{xl} - \mu_{kl})^2}{2\sigma^2_{kl}} \qquad (4)
\end{aligned}
$$

where ϕ_{xk} is the probability that the location x has been drawn by the topic k. The fourth term refers to the response variable g_x at location x and is drawn as a linear regression model over the topic assignment z_x with parameters $\{\eta, \delta^2\}$:

$$
\begin{aligned}
&E_q[log\ p(g_x|z_x, \eta, \delta^2)] = -\frac{1}{2} log(2\pi\delta^2) \\
&- \frac{1}{2\delta^2}(g_x^2 - 2g_x\eta^T\phi_x + \eta^T diag(\phi_x)\eta) \qquad (5)
\end{aligned}
$$

where ϕ_x is the vector of topic proportions ϕ_{xk} in the location x. Computing the derivatives of the KL divergence with respect to the parameters and setting them equal to zero, we obtain the update equations for the variational procedure. In particular, in the *variational E-step* we must update the variational parameters:

$$
\begin{aligned}
\phi_{xk} \propto &\frac{1}{\prod_{l=1}^{L} \sigma_{kl}} \exp\left[\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right) + \right. \\
&\left. \frac{g_x\eta_k}{\delta^2} - \frac{\eta_k^2}{2\delta^2} - \sum_{l=1}^{L} \frac{(f_{xl} - \mu_{kl})^2}{2\sigma^2_{kl}}\right] \qquad (6) \\
\gamma_k = &\alpha_k + \sum_{x=1}^{X} \phi_{xk} \qquad (7)
\end{aligned}
$$

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 $\Gamma_{1:L,1:K}, \eta, \delta^2$, in order to compute their optimal values. First, parameters μ_{kl} and σ^2_{kl} are computed for each feature l and topic k:

$$\mu_{kl} \propto \sum_{t=1}^{T}\sum_{x=1}^{X_t} \phi_{txk} f_{txkl} \qquad (8)$$

$$\sigma^2_{kl} \propto \sum_{t=1}^{T}\sum_{x=1}^{X_t} \phi_{txk}(f_{txkl} - \mu_{kl})^2 \qquad (9)$$

where it is worth noting that we have added the subindex t to indicate the frame number in the corpus.

Furthermore, during the training step, we use the ground-truth (GT) response value g_{tx} of all points in the corpus to learn the parameters of the linear regression model:

$$\eta_k \propto \frac{\sum_{t=1}^{T}\sum_{x=1}^{X_t} \phi_{txk} g_{tx}}{\sum_{t=1}^{T}\sum_{x=1}^{X_t} \phi_{txk}} \qquad (10)$$

$$\delta^2 = \frac{\sum_{t=1}^{T}\sum_{x=1}^{X_t} (g_{tx}^2 - 2g_{tx}\eta^T\phi_{ls} + \eta^T diag(\phi_{tx})\eta)}{\sum_{t=1}^{T} X_t} \qquad (11)$$

Training and predicting visual attention maps

In the learning phase, we get the optimal values of the parameters that maximize the ELBO of the log-likelihood. For that end, spatial locations of frames are non-uniformly sampled to ensure balanced datasets of salient and non-salient points. GT values for each spatial location $x \in X$ constitute also the responses g_x to be modeled as a linear regression model over topic proportions.

Furthermore, as the considered feature maps f_l have been designed to be proportional to the final visual attention, we aim to ensure that our model output grows with the feature maps (i.e., the larger a feature map the more likely the location attracts the visual attention). Hence, with the objective of learning subtasks that either attract or inhibit attention, we have designed topic-feature distributions as fixed-mean Gaussians: $\mu_{kl} = 0$ and $\mu_{kl} = 1$ for those topics inhibiting or attracting attention, respectively. This approach has led to superior performance than a model with free mean and variances.

Finally, once models are trained, in the test phase, attention is predicted at uniformly spaced locations x 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 linear regression over the topic assignments:

$$E[g_x|f_{x,1:L}, \alpha, \Gamma_{1:K}, \eta, \delta^2] \approx \eta^T E_q[z_x] = \eta^T \phi_x \qquad (12)$$

IV. EXPERIMENTS

A. Experimental Setup

The well-known freely-accessible CRCNS-ORIG [24] is used as benchmark dataset. The database contains eye movement recordings from eight distinct subjects watching 50

TABLE I

EVALUATION OF THE DIFFERENT NORMALIZATION STEPS, USING THE CONTEXT-GENERIC TOPIC APPROACH IN THE CRCNS-ORIG [24] DATABASE.

Normalization	[0, 1]	SFN	SFN + IF-IF
sAUC	$0,590$	$0,593$	$0,606$

different video clips (over 46,000 video frames, 25 minutes total, 640×480). Eye traces have been obtained using a 340 Hz ISCAN RK-464 eye-tracker. 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. For every frame, GT visual attention maps are obtained by convolving fixation density maps with a 2D Gaussian kernel ($\sigma = 8$).

In our experiments we will compare two approaches based on our probabilistic model: a) a *context-generic* model trained using frames belonging to videos in all the categories; and b) 7 *context-aware* models trained on those videos belonging to each category or genre. Finally, in order to test the performance over every video in the dataset, we have carried out a 4-fold cross validation procedure, so that at each iteration some videos were picked for evaluation. It is worth noting that all the frames of a video are always grouped together in the same set (train or test) to avoid over-fitting.

Many metrics have been proposed to compare visual attention models [25]. For our experiments, we have selected *Shuffled Area Under Curve (sAUC)*, which is one of the most commonly used. In the implementation used to estimate sAUC, which is detailed in [26], common fixation positions receive less credit for correct prediction so that center-bias effects of the spatial distribution of eye fixations are eliminated.

B. Results

Selecting the optimal number of topics

The main parameter in our model is the number K of subtasks or topics that contribute to model visual attention. For simplicity, we have used the same number of topics attracting ($\mu_{kl} = 1$) or inhibiting ($\mu_{kl} = 0$) visual attention. We have assessed our proposed *context-generic* topic model with respect to this parameter and found that the performance increases with the number of topics until $K = 30$, where it starts to remain more stable. Hence, we have selected an intermediate value $K = 40$ for the rest of the experiments.

Evaluation of normalization stages

In order to assess the performance of the normalization procedures, we have learned and tested our *context-generic* topic model with various alternative normalizations. In this experiment only the 10% of frames from the whole dataset were used. Results shown in Table I demonstrate that each proposed normalization helps to improve the system performance.

TABLE II

sAUC OBTAINED FOR THE PROPOSED CONTEXT-GENERIC AND CONTEXT-AWARE MODELS IN THE CRCNS-ORIG [24] DATABASE. RESULTS OF THE WELL-KNOWN GBVS [13] METHOD ARE ALSO INCLUDED FOR COMPARISON.

Context	Size	Frames	sAUC		
			C-Generic	C-Aware	GBVS [13]
Outdoor	17	$8,357$	$0,646$	$\mathbf{0,669}$	$0,586$
Videogames	9	$15,809$	$0,603$	$0,603$	$\mathbf{0,630}$
Commercials	4	$2,618$	$0,574$	$0,584$	$\mathbf{0,588}$
TV News	7	$8,071$	$0,539$	$\mathbf{0,554}$	$0,458$
Sports	5	$4,851$	$\mathbf{0,539}$	$0,537$	$0,511$
Talk Shows	4	$4,244$	$\mathbf{0,620}$	$0,599$	$0,537$
Others	4	$2,539$	$0,644$	$\mathbf{0,699}$	$0,691$
AVERAGE	50	$46,489$	$0,595$	$\mathbf{0,603}$	$0,572$

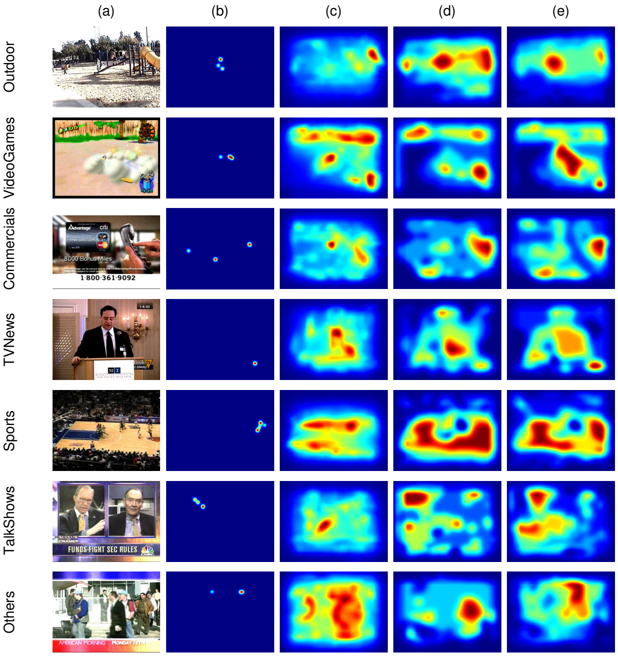

Fig. 3. Visual attention maps obtained for some example frames from CRCNS-ORIG [24] database. (a) Original frames. (b) GT visual attention maps. (c) GBVS: Harel et al. [13]. (d) Context-Generic. (e) Context-Aware.

A comparison between generic and context-aware visual attention models

Finally, we evaluate our method on different scenarios, learning both *context-generic* and *context-aware* models for the whole database and each category, respectively. Results obtained are provided in Table II. As can be seen, the context-aware models outperform the generic approach in most genres, which demonstrates that learning specific context-aware representations of visual attention over smaller training sets (the training videos belonging to each category) works better than generic models over larger datasets (including all video categories). Best scores are obtained for *Outdoor* and *Videogames* genres, due to the stronger influence assigned to motion-related features, which strongly help to distinguish moving objects. It is also noticeable the improvement achieved by the two alternatives with respect to the well-known GBVS [13], a reference BU method in the literature. Therefore, we can conclude saying that learning specific distributions for each video genre, we can notably enhace the system performance.

V. Discussion

In this paper, we have introduced a novel probabilistic topic model to estimate top-down visual attention in videos. In contrast to previous existing TD methods, our model incorporates a new intermediate level that decomposes the complex process of context-driven visual attention into a mixture of latent topics. Those topics represent subtasks that drive the visual attention and are unsupervisely inferred from data as combinations of distributions of low-level visual descriptors extracted from frames. In addition, by learning from human fixations, our proposal can be adapted to any particular scenario or task.

The experiments conducted in a dataset with videos organized by genre have shown how our approach successfully learns specific distributions for each category, thus enhancing the system performance with respect to a generic model for visual attention and outperforming a well known reference BU saliency method. However, it is still necessary to assess its capability to predict the more interesting task-driven visual attention, demonstrating its usefulness in end-user applications. For that end, further work will focus on the development of task-driven video datasets with human fixations to apply this approach.

Acknowledgment

This work has been partially funded by the Spanish Ministry of Education, Culture and Sports FPU programme and the National Grant TEC2014-53390-P of the Spanish Ministry of Science and Innovation.

References

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

[2] 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.

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

[4] V. Buso, I. Gonzlez-Daz, and J. Benois-Pineau, "Goal-oriented top-down probabilistic visual attention model for recognition of manipulated objects in egocentric videos," *Signal Processing: Image Communication*, vol. 39, Part B, pp. 418 – 431, 2015, recent Advances in Vision Modeling for Image and Video Processing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596515000892

[5] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1408–1424, July 2015.

[6] 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. [Online]. Available: http://dx.doi.org/10.1007/s12559-010-9094-8

[7] M. M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S. M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, March 2015.

[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. [Online]. Available: http://doi.acm.org/10.1145/641007.641116

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

[10] 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. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/3836989

[11] 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. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2012.89

[12] 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. [Online]. Available: http://dx.doi.org/10.1109/34.730558

[13] 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.

[14] 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.

[15] 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0042698910000052

[16] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 989–1005, June 2009.

[17] S. Han and N. Vasconcelos, "Biologically plausible saliency mechanisms improve feedforward object recognition," *Vision Research*, vol. 50, no. 22, pp. 2295 – 2307, 2010, mathematical Models of Visual Coding. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0042698910002786

[18] 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. [Online]. Available: + http://dx.doi.org/10.1167/8.7.32

[19] 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.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[21] 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. [Online]. Available: http://papers.nips.cc/paper/3328-supervised-topic-models.pdf

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

[23] 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.

[24] L. Itti and R. Carmi, "Eye-tracking data from human volunteers watching complex video stimuli," Dec 2009. [Online]. Available: http://crcns.org/data-sets/eye/eye-1/about

[25] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec 2013, pp. 1153–1160.

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