OBJECT RECOGNITION WITH TOP-DOWN VISUAL ATTENTION MODELING FOR BEHAVIORAL STUDIES

Vincent Buso, Iván González-Díaz, Jenny Benois-Pineau

OBJECT RECOGNITION WITH TOP-DOWN VISUAL ATTENTION MODELING FOR BEHAVIORAL STUDIES

Vincent Buso, Iván González-Díaz***

*University of Bordeaux
LaBRI
33405 Talence, France

*Jenny Benois-Pineau**

**Universidad Carlos III de Madrid
Dpt. of Signal Theory and Communications
Leganés, 28911, Madrid, Spain

ABSTRACT

Behavioural analysis in instrumental activities of daily living has become a powerful tool in clinical studies and rises the question of what objects are manipulated by patients. In this paper we present a top-down probabilistic visual attention model for manipulated object recognition in egocentric video content. Although arms often occlude objects and are usually seen as a burden for many vision systems, they become an asset in our approach, as we extract both global and local features describing their geometric layout and pose, as well as the objects being manipulated. We integrate this information in a probabilistic generative model, provide update equations that automatically compute the model parameters optimizing the likelihood of the data, and design a method to generate maps of visual attention that are later used in an object-recognition framework. This task-driven assessment reveals that the proposed method outperforms the state of the art in object recognition for egocentric video content.

***Index Terms*—** Saliency Maps; Object Recognition; Egocentric Vision; Vision Modelling; Image Processing; Video Processing

1. INTRODUCTION AND PREVIOUS WORK

Egocentric videos recorded by a mono or stereo camera worn by patients are more and more in focus nowadays since they provide a close-up view on actions and allow for an efficient analysis of objects manipulation [1]. Behavioural studies of subjects executing simple Instrumental Activities of Daily Living (IADL) are necessary in different clinical applications that is why the mobile of our research is the recognition of manipulated objects in egocentric video scenes for behavioural studies of patients with dementia. Indeed as studies show [2], the analysis of behaviour in IADL performance rises the question of what object is manipulated by the patient.

In our application scenario of egocentric videos recorded by a patient with a body worn camera, objects in the focus-of-attention of an observer analysing the visual scene (such as the medical doctor) show very clear correspondence with the objects being manipulated. However, the measurement of the subject fixations with eye-tracker devices is costly that is why the point of view of the medical doctor has to be modelled. Indeed, efficient ways of visual attention prediction have to be proposed for egocentric scenes.

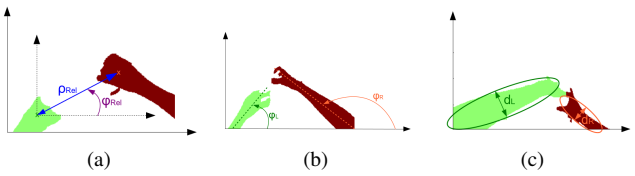
Since the pioneering work of Itti[3], modelling of visual attention for scene interpretation has become a very popular subject in image and video analysis. The so-called "predicted visual attention maps" used in feature selection and pooling [4, 5] in the problem of object recognition, have become a good competitor for heavy sliding window methods such as Deformable Part-Based Models (DPM)[6]. In all these and many other saliency approaches, the predicted visual attention maps are built accordingly to a "bottom-up" approach, which simulates sensitivity of Human Visual System(SVH) to colours, contrasts, orientation, residual motion[7],[8], and often incorporates the center-bias hypothesis, which means attraction of human gaze by the center of a frame [9]. In this paper on the contrary, we propose a top-down attention prediction knowing that the target of attention are manipulated objects.

The introduction of top-down factors into the classical bottom-up framework by extracting semantic clues (e.g., face, speech and music, camera, motion) was proven to provide impressive results [10, 11]. More recent works using machine learning approaches to learn top-down behaviours based on eye-fixation or annotated salient regions, have also proven to be very useful for static images [12, 13, 14] as well as videos [15, 16]. However, in practice, most videos contain many different attractors. Therefore, with such methods, it is impossible to generate a category-agnostic detector of the object of interest as they are based on a fixed set of pre-defined object categories.

In this paper we propose to use domain specific knowledge to present a new method for saliency maps computation based only on top-down components for egocentric video content. One of the main particularity of egocentric videos is the presence of hands and arms. Actors performing activities of daily living are indeed most-likely manipulating objects bare-handed. Hands are a commonly known burden in egocentric videos as they often obstruct objects resulting in a more challenging recognition process. However, our work consider hands and arms as assets for building top-down-only visual attention maps and without prior training of a fixed number of high semantic cues such as object categories [15, 16]. Thus, our saliency maps aim to measure the likelihood of pixels in the vicinity of hands to belong to a manipulated object. This function depends on the hands relative position with regard to the camera, their relative position with regard to each other, and the particular (learned) pose of manipulated objects under each hands configuration. This map is then used for psycho-visual weighting of frames signature in a supervised learning framework for object recognition.

In this work we completely remodel the approach firstly described in [17] with the following novel contributions:

- We totally avoid any segmentation of foreground objects, but instead require some manually annotated bounding boxes around objects of interest present in the training set.
- In contrast to our previous developments, in which the values of the model parameters were either set based on previous intuitions or cross-validated, here we propose an integrated data-driven learning framework that uses the Expectation-Maximization (EM) algorithm to automatically estimate parameters optimizing the likelihood of the data.



(a) (b) (c)

Fig. 1: Illustrations of the 6 global features. 1(a): *Relative location of hands*, 1(b): *Left and right arm orientations*, 1(c): *Left arm depth* and *Right arm depth with regard to the camera.*

The rest of the paper is organized as follows: in Section 2 we present Goal-Oriented Top-down visual attention model construction. In Section 3 we summarize the object recognition approach with predicted visual attention maps. Experiments, benchmarking and results are presented in Section 4. Section 5 concludes the paper and proposes perspectives of this work.

2. GOAL-ORIENTED TOP-DOWN VISUAL ATTENTION MODEL

In this section we present our model of visual attention prediction for the task of manipulated object recognition. It relies on extraction of the arms/hands automatically fulfilled for each frame using the approach introduced by Fathi et al. [18].

2.1. Defining global and local features

We propose to build our attention model as a combination of two distinct sets of features. The first one is a set of global features that describe the geometric configuration of the segmented arms. The features are then clustered into a pre-defined number of states/configurations. Prior knowledge on saliency is then generated for each of these global states. The second one is based on local features which will allow to adaptively relocate these saliency priors.

We introduce global features based on the geometry of arms in the camera field of view correlated with manipulated object size and position. Each arm, from elbow to the hand extremity, is approximated by an elliptic region in the image plane. Hence an ellipse is first fitted to each segmented arm area and, then, the six following global features are extracted:

- *Relative location of hands*: The magnitude ρ_{Rel} and phase φ_{Rel} between hands centres are extracted (see Fig. 1(a)) since they are strong indicators of the objects width and holding pose, respectively.
- *Left arm orientation* and *Right arm orientation*: As illustrated on Fig. 1(b) the orientation of each arm (φ_L and φ_R) is extracted. The arms are mostly oriented depending on the objects being manipulated.
- *Left arm depth* and *Right arm depth*: an object size is correlated with the "depth" of the arms. In this work, the body-worn cameras do not provide a real depth information. A trivial approximate of the "depth" of an arm is the minor axis length d_L and d_R) of the fitted ellipse (see figure 1(c)).

A vector $\mathbf{g} = (\rho_{Rel}, \varphi_{Rel}, \varphi_L, \varphi_R, d_L, d_R)^T$ containing these six geometrical features is computed for each image in the training set. Then, the whole set is clustered into K global appearance models $z_k, k = 1..K$ using k-means [19]. The aim of this pre-processing stage is to use the models to initialize our probabilistic approach (to be described in the next section). It is worth noting that a Z-score normalization has been performed over the data, in order to prevent out-weighting features with large range over attributes with small ones [20].

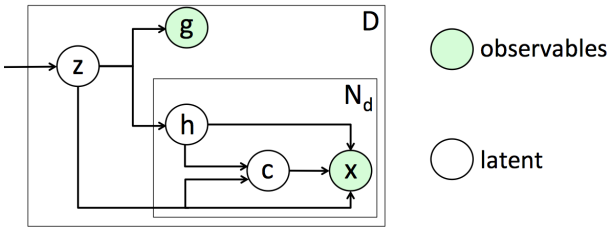


Fig. 2: Graphical model of our approach Top-down visual attention modelling with manipulated objects. Nodes represent random variables (observed-shaded, latent-unshaded), edges show dependencies among variables, and boxes refer to different instances of the same variable.

Furthermore, we consider some "local" features that help to compute a saliency distribution given by the global arm configuration of a frame. These features are the coordinates of hand centres $\mathbf{c}$, the hand indicator h (left or right), and the candidate pixels $\mathbf{x}$ around the hand to belong to the object being manipulated.

2.2. A Probabilistic Model for Top-down Visual Attention Prediction

As a human observer would be attracted by hand-manipulated objects, we consider the joint locations of arms/hands and objects as predictors of top-down visual attention. Hence our probabilistic model for top-down visual attention incorporates distributions of both global and local features presented in the previous section. The graphical model of our approach is shown in Fig. 2. Based on this, given a corpus of D training images our objective is, for each image d, to learn the process that chooses a set of N_d salient spatial locations $\mathbf{x}$.

To do so, the generative process first randomly picks a global arm model $\mathbf{z}_k$ from the K candidates. K corresponds to the number of clusters as defined in section 2.1, and remains an open parameter in our model. Then, depending on the selected global model $\mathbf{z}_k$, global ($\mathbf{g}$) and local ($\mathbf{x}, \mathbf{c}, h$) features are drawn from the particular conditional distributions $p(\mathbf{g}|\mathbf{z}_k)$ and $p(\mathbf{x}, \mathbf{c}, h|\mathbf{z}_k)$, respectively. Here, h is an index variable with two possible values $h = 0, 1$ for left and right hands, respectively

In the following paragraphs we will first introduce the distributions modelling both global and local features. Then we will combine them together to build the expression of the corpus likelihood. Next, we will describe the learning process and the update equations that allow us to obtain the optimal model parameters maximizing this likelihood. Finally, we will describe how this model is used to generate saliency maps from data.

Distributions of Global Features: We define the conditional distribution that models the global features given the component z_k with a Gaussian pdf $p(\mathbf{g}|z_k) = \mathcal{N}(\mathbf{g}; \mu_k^g, \Sigma_k^g)$, with mean vector μ_k^g and covariance matrix Σ_k^g.

Distributions of Local Features: Concerning the local features, for each elementary arms model z_k, we draw N_d points at spatial locations $\mathbf{x}$ considered as salient. For that end, we start by picking a hand (left or right) following the distribution $p(h|z_k)$. Next, once the hand is chosen, we randomly locate its centre by drawing its

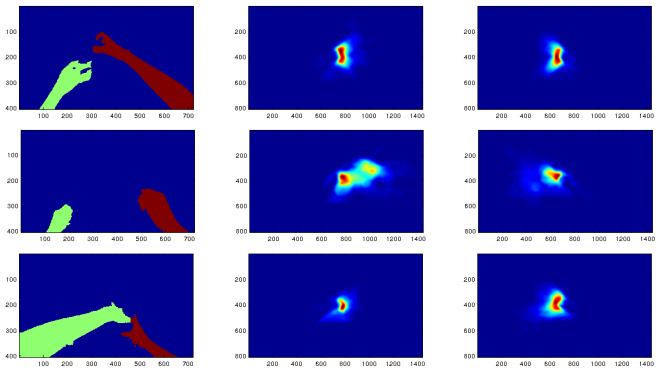


Fig. 3: Two examples of the obtained experimental distributions $p(\mathbf{x}_i|h_j, \mathbf{c}, z_k)$. Left column: arms segmentation representing the global model. Middle column: left hand distribution. Right column: right hand distribution.

coordinates $\mathbf{c}$ using the distribution $p(\mathbf{c}|h, z_k)$. Finally, we use the conditional distribution $p(\mathbf{x}|h, \mathbf{c}, z_k)$ to randomly choose a spatial location $\mathbf{x}$ that belongs to the object being manipulated. This distribution models the probability of a pixel to belong to the object being manipulated given the current geometric configuration of arms and hands.

Putting everything together and marginalizing over the variable h, we can expand the distribution involving the *local features*:

$$p(\mathbf{x}, \mathbf{c}, h|z_k) = \sum_{j=0}^{1} p(h_j|z_k)p(\mathbf{c}|h_j, z_k)p(\mathbf{x}|h_j, \mathbf{c}, z_k) \quad (1)$$

Now, we can define the particular conditional distributions that model each variable:

1. The selected hand is given by a discrete distribution $p(h_j|z_k) = \alpha_{jk}$, with $\sum_{j=0}^{1} \alpha_{jk} = 1$.
2. The hand centre $\mathbf{c}$ follows a Gaussian distribution $p(\mathbf{c}|h_j, z_k) = \mathcal{N}(\mathbf{c}; \mu^c_{jk}, \Sigma^c_{jk})$.
3. The spatial location $\mathbf{x}$ is defined with an experimental discrete distribution: $p(\mathbf{x}_i|h_j, \mathbf{c}, z_k) = \beta_{kji}$, so that $\sum_{i=0}^{L^2} \beta_{kji} = 1$. This distribution is defined over a square 2D box of size LxL centred at $\mathbf{c}$ built by superimposing all accordingly-centred annotated objects from images belonging to the cluster z_k. In Fig. 3 we show some empirical examples of this distribution.

Log-Likelihood: From the graph depicted in Fig. 2, the likelihood of the corpus given the model parameters: $\theta = \{\pi, \mu^g, \Sigma^g, \alpha, \mu^c, \Sigma^c, \beta\}$ can be defined by means of a mixture of K components:

$$\mathcal{L} = p(\mathbf{x}, \mathbf{g}|\theta) = \prod_{d=1}^{D} p(\mathbf{z}_d)p(\mathbf{g}_d|\mathbf{z}_d) \prod_{i=1}^{N_d} p(\mathbf{x}_i, \mathbf{c}_i, h_i|\mathbf{z}_d) \quad (2)$$

where $p(z_k) = \pi_k$ is discrete with parameter π_k and stands for the prior distribution of the global arm models (weights of components in the mixture).

If we marginalize over the latent arm models we get:

$$\mathcal{L} = \prod_{d=1}^{D} \sum_{k=1}^{K} p(z_k)p(\mathbf{g}_d|z_k) \prod_{i=1}^{N_d} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z_k) \quad (3)$$

Taking logarithms and applying the Jensen's inequality one can obtain a lower bound of the log-likelihood:

$$\log \mathcal{L} \geq \sum_{d,k}^{D,K} \phi_{dk} \left[\log \left(p(z_k)p(\mathbf{g}_d|z_k) \cdot \prod_{i=1}^{N_d} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z_k) \right) - \log \phi_{dk} \right] \quad (4)$$

where we have introduced a new variable $\phi_{dk} = p(z_k|\mathbf{g}_d, \mathbf{x})$ which stands for the posterior distribution of the arms model given the observed variables and obeys $\sum_k \phi_{dk} = 1$.

In addition, we can also lower-bound the term of log-likelihood related to the local features if we apply again the Jensen's inequality:

$$\log \mathcal{L}_{local} = \sum_{dki} \phi_{dk} \log \sum_{j=0}^{1} p(h_j|z_k)p(\mathbf{c}|h_j, z_k)p(\mathbf{x}_i|h_j, \mathbf{c}, z_k)$$

$$\geq \sum_{dkij} \gamma_{dkij} \left[\log p(h_j|z_k)p(\mathbf{c}|h_j, z_k)p(\mathbf{x}_i|h_j, \mathbf{c}, z_k) - \log \gamma_{dkij} \right] \quad (5)$$

where $\gamma_{dkij} = p(h_j|z_k, \mathbf{c}, \mathbf{x}_i)$ is the posterior distribution of the selected hand once the global model, the center and the spatial location are known.

Inference: We aim to learn the set of optimal model parameters $\theta = \{\pi, \mu^g, \Sigma^g, \alpha, \mu^c, \Sigma^c, \beta\}$ that maximize the log-likelihood. For that end, we have used the Expectation-Maximization (EM). Due to the lack of space, we omit the algebra to obtain the EM update equations.

In the *E-Step*, the algorithm computes the expected values of the posterior distributions ϕ_{dk}, γ_{dkij}:

$$\phi_{dk} \propto p(z_k)p(\mathbf{g}_d|z_k) \prod_{i=1}^{N_d} p(\mathbf{x}_i, \mathbf{c}_i, h_i|z_k) \quad (6)$$

$$\gamma_{dkij} \propto p(h_j|z_k)p(\mathbf{c}|h_j, z_k)p(\mathbf{x}_i|h_j, \mathbf{c}, z_k) \quad (7)$$

In the *M-Step*, our algorithm updates the values of the model parameters:

$$\pi_k = \frac{1}{D} \sum_d \phi_{dk} \quad (8)$$

$$\mu^g_k \propto \sum_d \phi_{dk}\mathbf{g}_d \quad (9)$$

$$\Sigma^g_k \propto \sum_d \phi_{dk}(\mathbf{g}_d - \mu^g_k)(\mathbf{g}_d - \mu^g_k)^T \quad (10)$$

$$\alpha_{jk} \propto \sum_{di} \phi_{dk}\gamma_{dkij} \quad (11)$$

$$\mu^c_{jk} \propto \sum_d \phi_{dk}c_{dj} \sum_i \gamma_{dkij} \quad (12)$$

$$\Sigma^c_{jk} \propto \sum_d \phi_{dk}(c_{dj} - \mu^c_{jk})(c_{dj} - \mu^c_{jk})^T \sum_i \gamma_{dkij} \quad (13)$$

$$\beta_{kji} \propto \sum_d \phi_{dk}\gamma_{dkij} \quad (14)$$

Building Saliency Maps: Once the optimal parameters have been learned, we can build a saliency map by measuring the saliency of every pixel location. For that end, the *saliency value of a pixel* $S(\mathbf{x})$ can be defined as its likelihood over the proposed generative model for saliency $S(\mathbf{x}) = p(\mathbf{x}_i, \mathbf{g}|\theta)$

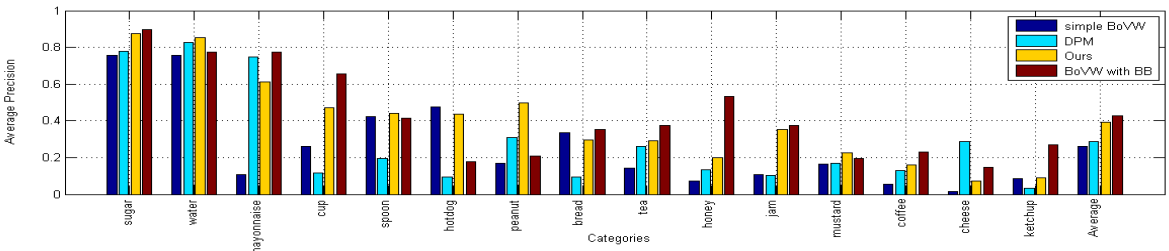


Fig. 4: Object recognition performances between different paradigms. The results are given in average precision per category and averaged.

3. OBJECT RECOGNITION APPROACH

In our previous work [21] we have proposed an object recognition approach in the family of methods which use psycho-visual weighting [22] of the conventional Bag-of-Visual-Words (BoVW)[23],[24] paradigm. Here we compute SURF descriptors[25] on a dense grid, then a visual dictionary is built from all descriptors extracted in the training video database. When computing the signatures, the contribution of each descriptor to the histogram is weighted by the maximal predicted saliency value in a small circular patch around it. Once each image is represented by its weighted histogram of visual words, an SVM classifier [26] is used with χ^2 kernel.

4. EXPERIMENTS AND RESULTS

In the context of this study, developing a model of top-down visual saliency aims at improving object recognition performances in patient-worn egocentric videos. The proposed approach is compared to other well-known paradigms for object recognition applied to this video content.

4.1. Dataset and setup

The experiments were conducted on the GTEA dataset [18] since it is a publicly available database of egocentric videos of 4 subjects performing 7 types of instrumental activities of daily living. The segmentations of arms are provided for 17 videos, a subset of which is used for training our distributions. In [18] the frames were annotated with the objects of interest but we manually extend this annotation by drawing bounding boxes on them. The dictionary size for computing the visual dictionary in BoVW was fixed at 4000. Finally, we set the number of global configurations K to the one which gave the maximum recognition performances with cross validation over the training set, that is to say $K = 70$.

4.2. Object recognition performances

In Fig. 4 we present a comparison of object recognition performances between our approach (denoted as 'Ours'), the reference BoVW scheme, and a well-known technique considered as State-of-the-art in egocentric vision: the discriminatively-trained Deformable Part Model (DPM) [6], a sliding window technique that has been taken as the object detection paradigm by the authors of the ADL dataset [2]. Our method outperforms these two famous paradigms for object recognition by achieving absolute improvements of mean Average Precision (mAP) of 13% over BoVW and 10.7% over DPM. Here the "ideal" case BoVW with BB is added for the upper bound estimate. It is a BoVW scheme where descriptors were extracted only in manually annotated bounding boxes around objects of interest. We consider these bounding boxes as "ideal" saliency maps. As can be seen from the mAP score (last set of bars in figure 4, reported in table 1 for clarity), our method not only outperforms any other in this kind of video content but also achieves very close performances to the "ideal" case. All the improvements were backed up by performing Student's t-tests with significance level of 0.05.

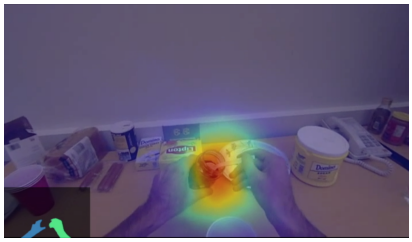


Fig. 5: Illustration of saliency maps obtained with our model as a heat map in alpha-blending with the original frame. The The top left corner shows the corresponding arm segmentation.

	Simple BoVW	DPM	Ours	BoVW with BB
mAP	0.262	0.285	**0.392**	0.425

Table 1: Mean Average Precision (mAP) Results over all categories for all considered object recognition paradigms

5. CONCLUSIONS AND PERSPECTIVES

In this paper we have presented a top-down task-driven visual attention model for the goal of object recognition in egocentric videos. In this kind of video content we aim at facilitating behavioural studies of activities of daily living for clinical scenarios. However we believe this model of saliency could extend to any scenario willing to detect manipulated objects in egocentric videos, especially with the current development of wearable cameras such as GoPro. Our method only requires prior input from a hand detector [27], and annotations of objects in the training set, after that it becomes fully automatic and, as presented in section 4.2, outperforms state of the art results in the GTEA egocentric dataset (close to what we defined as the "ideal" method requiring user inputs). Our main perspective now is to make this visual attention model auto-dependent by adapting and incorporating a state of the art Hand detection scheme in order to review its performances on different publicly available egocentric datasets.

6. ACKNOWLEDGEMENTS

This research is supported by the EU FP7 PI Dem@Care project #288199.

7. REFERENCES

[1] Ali Borj and Laurent Itti, "Stereovision and augmented reality for closed-loop control of grasping in hand prostheses," *J.Neural Eng*, vol. 35, no. 11, pp. 1–17, November 2014.

[2] Hamed Pirsiavash and Deva Ramanan, "Detecting activities of daily living in first-person camera views," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*. IEEE, 2012.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 20, no. 11, pp. 1254–1259, 1998.

[4] Alireza Fathi, Yin Li, and James M. Rehg, "Learning to recognize daily actions using gaze," in *Proceedings of the 12th European conference on Computer Vision - Volume Part I*. 2012, ECCV'12, pp. 314–327, Springer-Verlag.

[5] K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato, "Coupling eye-motion and ego-motion features for first-person activity recognition.," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,2012*. 2012, pp. 1–7, IEEE.

[6] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analyisis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[7] O. Brouard, V. Ricordel, and D. Barba, "Cartes de Saillance Spatio-Temporelle basées Contrastes de Couleur et Mouvement Relatif," in *Compression et representation des signaux audiovisuels, CORESA 2009*, March 2009, p. 6 pages.

[8] Ali Borj and Laurent Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. on Pattern Analysis and Machine INtelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[9] Benjamin W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14, pp. 1–17, Nov. 2007.

[10] Y. F. Ma, X. S. Hua, L. Lu, and H. Zhang, "A generic framework of user attention model and its application in video summarization.," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[11] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection.," in *NIPS*, John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, Eds. 2007, Curran Associates, Inc.

[12] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 989–1005, 2009.

[13] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "Sun: Top-down saliency using natural statistics," 2009.

[14] A. Torralba, M. S. Castelhano, A. Oliva, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search," *Psychological Review*, vol. 113, pp. 2006, 2006.

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

[16] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," *Int. J. Comput. Vision*, vol. 90, no. 2, pp. 150–165, Nov. 2010.

[17] V. Buso, I. Gonzalez-Diaz, and J. Benois-Pineau, "Goal-oriented top-down probabilistic visual attention model for recognition of manipulated objects in egocentric videos," *Signal Processing: Image Communication [Submitted]*, vol. Recent Advances in VM4IVP, 2014.

[18] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*. 2011, pp. 3281–3288, IEEE.

[19] S. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Inf. Theor.*, vol. 28, no. 2, pp. 129–137, Sept. 2006.

[20] L. Shalabi Al and Z. Shaaban, "Normalization as a preprocessing engine for data mining and the approach of preference matrix," in *Proceedings of the International Conference on Dependability of Computer Systems*. 2006, DEPCOS-RELCOMEX '06, pp. 207–214, IEEE Computer Society.

[21] I. González Díaz, V. Buso, J. Benois-Pineau, G. Bourmaud, and R. Megret, "Modeling instrumental activities of daily living in egocentric vision as sequences of active objects and context for alzheimer disease research," in *Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare*. 2013, MIIRH '13, pp. 11–14, ACM.

[22] R. de Carvalho Soares, I.R. da Silva, and D. Guliato, "Spatial locality weighting of features using saliency map with a bag-of-visual-words approach," in *IEEE 24th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2012, vol. 1, pp. 1070–1075.

[23] J. Sivic and A. Zisserman, "Video google : A text retrieval approach to object matching in videos," in *Proceedings of the International Conference on Computer Vision*, Oct. 2003, vol. 2, pp. 1470–1477.

[24] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[25] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, June 2008.

[26] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[27] Cheng Li and Kris M. Kitani, "Pixel-Level Hand Detection in Ego-centric Videos," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3570–3577, June 2013.