Buso, V., González-Díaz, I. & Benois-Pineau, J. (2015). 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.

DOI: 10.1016/j.image.2015.05.006

Goal-oriented top-down probabilistic visual attention model for recognition of manipulated objects in egocentric videos

Vincent Buso[a], Iván González-Díaz[b], Jenny Benois-Pineau[a]

[a]*Laboratoire Bordelais de Recherche en Informatique, Université Bordeaux, 33405 Talence, France.*
[b]*Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés, 28911, Madrid, Spain.*

Abstract

We propose a new top down probabilistic saliency model for egocentric video content. It aims to predict top-down visual attention maps focused on manipulated objects, that are then used for psycho-visual weighting of features in the problem of manipulated object recognition. The model is probabilistically defined using both global and local appearance features extracted from automatically segmented arm areas and objects. A psycho-visual experiment has been conducted in a guided framework that compares our proposal and other popular state-of-the-art models with respect to human gaze fixations. The obtained results show that our approach outperforms several popular bottom-up saliency approaches in a well-known egocentric dataset. Furthermore, an additional task-driven assessment for object recognition in egocentric video reveals that the proposed method improves the performance of several state-of-the-art techniques for object detection.

Keywords: Saliency Maps; Egocentric Vision; Object Recognition; Vision Modeling; Image Processing; Video Processing

Email addresses: vbuso@labri.fr (Vincent Buso), igonzalez@tsc.uc3m.es (Iván González-Díaz), benois-p@labri.fr (Jenny Benois-Pineau)

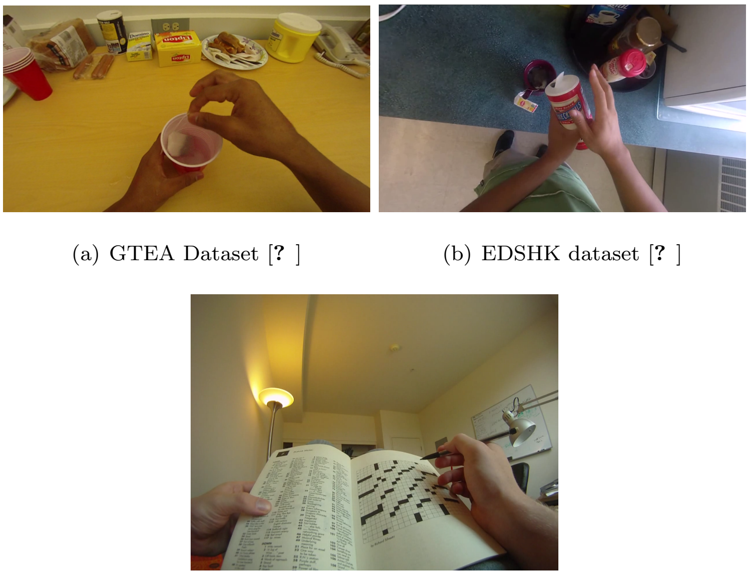
1. Introduction and motivation

The rationale and application of this research is in the objective assessment and life-logging of Alzheimer patients in their Instrumental Activities of Daily Living (IADLs). For this particular task, egocentric video analysis has gained a lot of interest. Indeed this kind of video content is recorded by cameras worn by a person, representing a cheap and effective way to record users' activity, and offering a unique point of view on the manipulated objects (see Figure 1). Recent studies have demonstrated how crucial is the recognition of manipulated objects for activity recognition under this scenario [? ?]. Considering methods for object recognition, two kinds of approaches can be identified in the literature: those relying on sliding windows, and those ones that first try to segment the foreground area containing the object of interest.

Concerning the first type, the authors of [?] applied the well-known Discriminatively Trained Deformable Part-Based (DPM) Models [?] to egocentric video. The second kind of approaches follows the well-known paradigm of foreground object segmentation to guide the object recognition process. The authors of [?] proposed a method that firstly segments the foreground areas from the background of each frame. Once the segmentation is made, the method detects and labels regions associated with the hands and the object being manipulated, respectively, and finally assigns an object label to the frame.

Concerning the second line of research, to drive the recognition process to relevant areas in the images, saliency or visual attention modelling was incorporated to the object recognition paradigms, showing an increase in the system performances [? ?]. In this paper, we will follow this line of research, as we are particularly interested in the modelling of human visual attention based on task-oriented top-down cues.

Generally speaking, two types of attention are commonly distinguished in the literature: bottom-up or stimulus-driven and top-down attention or goal-driven. [?],[?]. The authors of [?] define the top-down attention as the voluntary allocation of attention to certain features, objects, or regions in space. They also



(a) GTEA Dataset [?]

(b) EDSHK dataset [?]

(c) ADL dataset [?]

Figure 1: Examples of egocentric datasets illustrating the unique point of view on manipulated objects.

state that attention is not only voluntary directed as low-level salient stimuli can also attract attention, even though the subject had no intention to attend these stimuli. A recent study [?] about how saliency maps are created in the human brain, shows that an object captures our attention depending both on its bottom-up saliency and top-down control.

Modelling of human visual attention has been an intensively explored research subject since the last quarter of the 20th century and nowadays the majority of saliency computation methods are designed from a bottom-up perspective [?]. Bottom-up models are stimulus-driven, mainly based on low-level properties of the scene such as color, gradients orientation, motion or even depth. Consequently, bottom-up attention is fast, involuntary and, most likely feed-forward [?].

One of the first complete models of visual attention was proposed as a fusion of features based on the Human Visual System modelling (HVS) by Itti [?].

Since then, much work has been made in this domain ([? ? ?]). The reader is referred to a recent benchmark of several saliency models for more details [?].

However, although the literature concerning models of top-down attention is clearly less extensive, the introduction of top-down factors (e.g., face, speech and music, camera motion) into the modelling of visual attention has provided impressive results in previous works [? ?]. In addition, some attempts in the literature have been made to model both kinds of attention for scene understanding in a rather "generic" way. In [?] the authors claim that the top-down factor can be well explained by the focus in image, as the producer of visual content always focuses his camera on the object of interest. Nevertheless, it is difficult to admit this hypothesis for expressing the top-down attention of the observer of the content: it is always task-driven [?].

More recent works using machine learning approaches to learn top-down behaviours based on eye-fixation or annotated salient regions, have proven also to be very useful for static images [? ? ?] as well as for videos [? ?]. Furthermore, with advent of Deep Learning Networks (DNN), some novel approaches have been designed in the field of object recognition, which build class-agnostic object detectors to generate candidate salient bounding-boxes which are then labeled by later class-specific object classifiers [? ?]. However, it seems impossible for us to propose a universal method for prediction of the top-down visual attention component, as it is voluntary directed attention and therefore it is specific for the task of each visual search. Nevertheless, the prior knowledge about the task the observer is supposed to perform, allows extracting semantic clues from the video content which would ease such a prediction.

The current state-of the art in computer vision allows detection of some categories of objects with a high confidence. A variety of face or skin detectors have been proposed since the last two decades [?]. Hence, when modelling a top-down attention in a specific visual search task, we can use such "easily recognisable" semantic elements that are relevant to the specific task of the observer and may help to identify the real areas/objects of interest.

In this paper we propose to use domain specific knowledge to predict top-down visual attention in the task of recognizing manipulated objects in egocentric video content. In particular, our "recognisable elements" that are relevant to the task, are the arms and hands of the user wearing the camera and performing the action. Their quantized poses with regard to different elementary components of a complex action such as object manipulation will help in the definition of the area where the attention of the observer searching for manipulated objects will be directed. We evaluate our model from two points of view: i) prediction strength of gaze fixations of subjects observing the content with the goal of recognition of a manipulated object, and ii) performance in the target object recognition by a machine learning approach.

The rest of the paper is organized as follows: in section 2 we present our approach to generate top-down visual saliency maps. Section 3 describes the dataset, different experimental set-ups and provides the evaluation of the results. Finally section 4 draws main conclusions of this work and introduces research perspectives.

2. Goal-oriented top-down visual attention model

In this section we define a model of visual attention prediction in the task of manipulated object recognition. Our model relies on the detection and segmentation of some objects, considered as references, that help to locate the real areas of interest in a scene, namely the objects being manipulated. In our proposal, arms/hands are automatically computed for each frame using the approach introduced by Fathi et al. [**?**].

We propose to build our model as a combination of two distinct sets of features: global and local. The former describes the geometric configuration of the segmented arms, which are clustered into a pre-defined set of states/configurations. This global information is used to select one of the components in a mixture model. The second set, concerning the local features, is then modelled using the particular distributions corresponding to the selected global component.

Since the original approach in [**?**] generates not only an automatic segmentation masks for arms/hands, but also for manipulated objects, one can wonder if building saliency maps around manipulated objects is needed as the objects have been already segmented. The automatic object segmentation tends to supply very small parts of objects due to the occlusions by hands (see Figure 9(c)). Our intended saliency maps, on the contrary, cover the whole zone of interest even in the presence of occlusions. As we will show in the experimental section, even fitting a 2D Gaussian on the segmented objects as a trivial approximate of salient zones, in order to unify the possible segmented parts of objects, provide poor object recognition results.

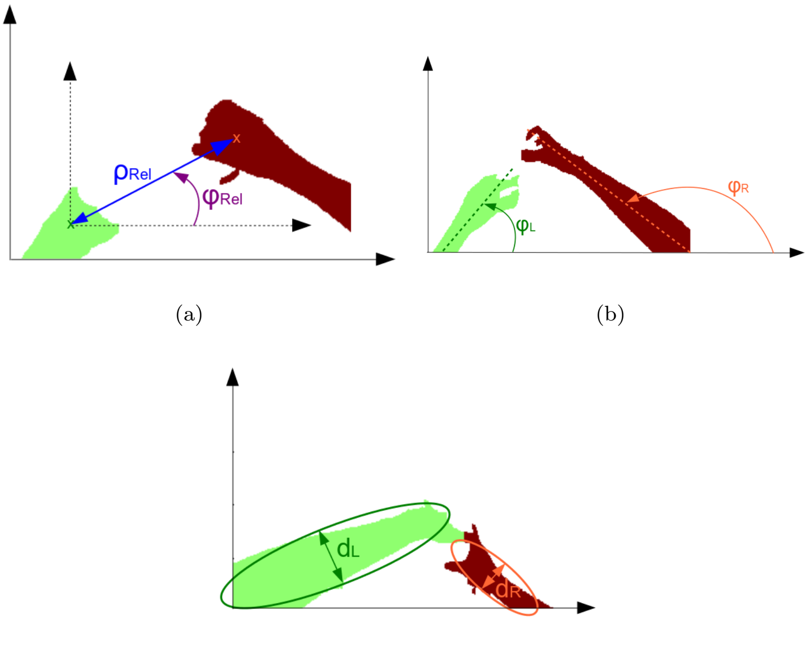
2.1. Defining global features

The features we propose are based on the geometry of arms in the camera view field, which is 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, several global features are defined, namely:

- *Relative location of hands*: Two features are extracted that encode the relative location of one hand with respect to the other (see figure 2(a)). For that end, taking the left hand centre as the origin of coordinates, the vector that joins the origin and the right hand is represented by means of its magnitude ρ_{Rel} and phase φ_{Rel} . Magnitude and phase are strong indicators of the objects width and holding pose, respectively.

- *Left arm orientation* and *Right arm orientation*: As illustrated on figure 2(b) the orientation of each arm (φ_L and φ_R) is defined by the angle between principle axis of ellipse and Y-axis in image plane. The arms are mostly oriented depending on the objects being manipulated, e.g.: holding a cup or pouring something (milk, juice, ...) present usually distinguishable arms orientations.

- *Left arm depth* and *Right arm depth with regard to the camera*: an object size is likely to be correlated with the "depth" of the arms, i.e. a measure



(a)

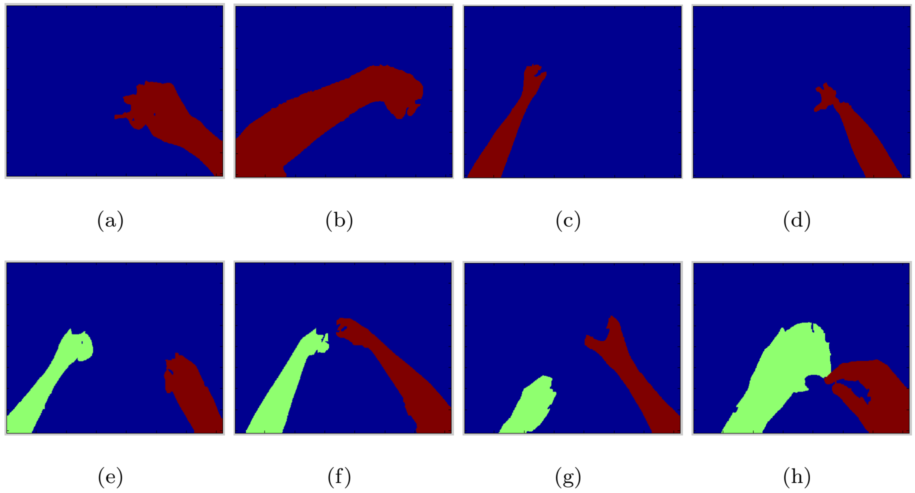
(b)

(c)

Figure 2: Illustrations of the 6 global features. 2(a): *Relative location of hands*, 2(b): *Left arm orientation*, 2(c): *Left arm depth* and *Right arm depth with regard to the camera.*

of its closeness to the camera. 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 2(c)).

A vector $\mathbf{g} = (\rho_{Rel}, \varphi_{Rel}, \varphi_L, \varphi_R, d_L, d_R)$ containing these six geometrical features is computed for each image in the training set, and then clustered into K global appearance models using k-means algorithm. It is worth noting that a Z-score normalization has been performed over the data, in order to prevent outweighing features with large range over attributes with small ones [**?**]. Figure 3 illustrates results in case of 8 clusters in our training dataset. The difference between the global appearance states (a) - (h) is easily noticeable.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 3: Representation of the arm segmentations closest to the centre of 8 global appearance model clusters. Each cluster is represented by the sample that is closest to the cluster centre.

2.2. Defining local features

Global appearance models define the most common states in which the arms can be found. Depending on these models, the zones of interest are different and the saliency computation needs to be adapted to. The "local" features we introduce serve for refining the underlying saliency distribution in the frame for a given global state. These features are the coordinates of a hand centre **c** (or hand centres in case the global state contains two hands). Their computation is also based on geometrical considerations.

Intuitively, when only the hand appears in the image, the hand centre **c** should be situated around the barycentre of the whole segmented image. Similarly, if the whole arm appears such on figure 4, the hand centre should be located closer from the extremity of the arm. Looking at Fig. 4, let us define two segments: x_{hs} is the segment that joins the beginning of the arm (origin of coordinates) with the beginning of the hand, and x_{ae} is the full arm-length. We have observed that the ratio $d = x_{hs}/x_{ae}$ is closely related with the ratio r between the minor and major axis of the fitted ellipse. In particular, to establish this relationship, we have randomly select 2615 arm segmentations for

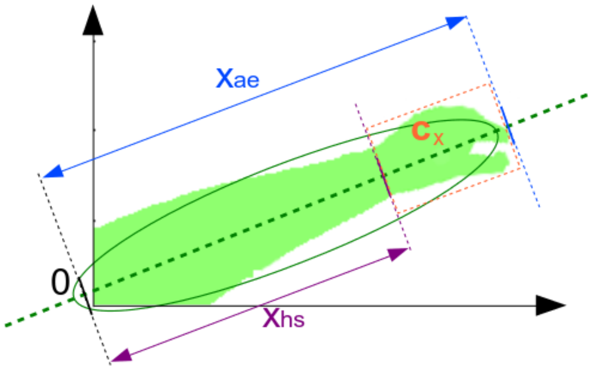


Figure 4: Illustration of the hand centre **c** computed as the barycentre of the orange box and the key points around: x_{hs} the starting position of the hand on the major ellipse axis, and x_{ae} the end position of the whole arm.

which we annotated the hands starting points x_{hs} over the major axis of the ellipse (represented as blue dots in Fig. 5), and then optimized an exponential model as:

$$d(r) = ae^{br} \tag{1}$$

where a and b are the coefficients computed by direct exponential fitting, r is the ratio between minor/major axis of the ellipse fitting the segmented arm, and $d(r)$ gives the ratio between the starting points x_{hs} of hands on the major ellipse axis and the arm length x_{ae} as a function of r. The results of this optimization are shown in Fig. 5) (continuous red line). Finally, the 2 dimensional center **c** coordinates are then defined as the barycentre of the segmented area that lies between the starting point of the hand and the end of the arm (the barycentre of the orange dotted box in Fig. 4). Computed “hand centre” **c** coordinates will serve to generate the resulting hand-related saliency map.

In order to evaluate the robustness against outliers we also performed regression using RANSAC [**?**] (with two thresholds, see green dashed and cyan dotted lines in Fig. 5). For all annotated data we computed the hand centres as well as the ones obtained with our three regression methods. Table 1 shows the

	Direct exponential regression	Exponential regression with Ransac, Threshold=0.1	Exponential regression with RANSAC, Threshold=0.01
Average distance (in pixels)	18.6	17.9	**17.5**

Table 1: NSS mean scores (with standard deviations) between human fixation points and different saliency map models. Our model outperforms the others

absolute average distances in pixels between the centres annotated by human annotators and the ones obtained with our three methods.

These results show that even though there are outliers, the final center approximations are very close (the difference is of order of 1 pixel) with and without outlier rejection by RANSAC. Furthermore, we noticed that stronger errors occur when the size of the hand is large (close-up view) in which case the saliency map are also larger.

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

As a human observer would be attracted to the objects manipulated by hands, we consider the joint locations of arms/hands and objects as predictors of top-down visual attention. Hence, we have developed a probabilistic model for top-down visual attention that incorporates both global and local features distributions. The graphical model of our approach for Top-Down visual attention is shown in Fig. 6. Given a corpus of D training images, the objective is to learn the process that chooses a set of N salient spatial locations $\mathbf{x}$ within each frame.

Let us first introduce a simplified model considering just the set of K global arm configurations $\mathbf{z} = \{z_1, ..., z_K\}$, and their relationship with the global features $\mathbf{g}$. Given $\mathbf{z}$, the probability density function (pdf) of the vector $\mathbf{g}$ can be

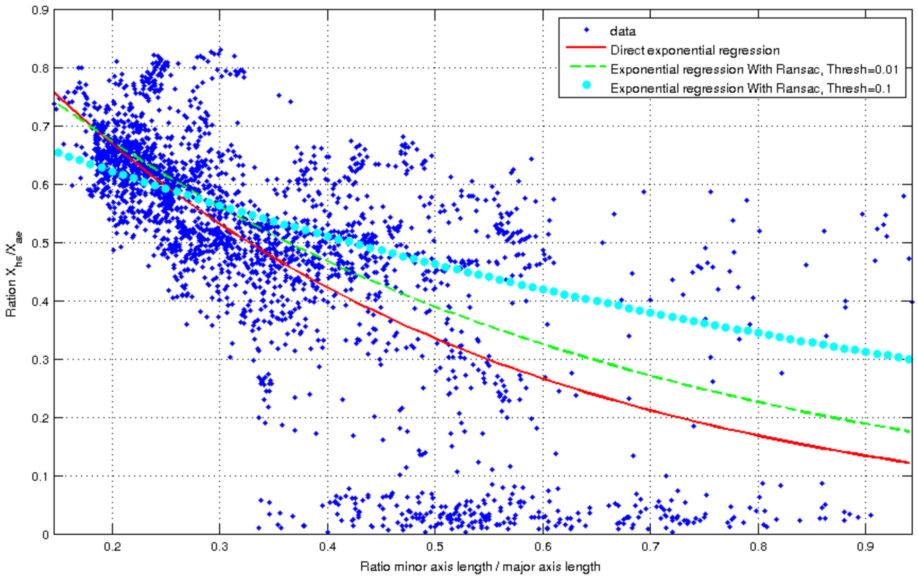


Figure 5: Graph representing the ratio between hands beginning and arm length depending on the minor/major axis lengths of the ellipse fitting the segmented arms. Blue dots correspond to the values manually annotated, red continuous line to the direct exponential regression, dotted cyan and dashed green line to the exponential regressions with RANSAC and different thresholds

modelled as a Gaussian mixture.

$$p(\mathbf{g}|\mathbf{z}) = \sum_{k=1}^{K} w_k p(\mathbf{g}|z_k) \tag{2}$$

Here K is the number of clusters in section 2.1, and remains an open parameter in our model. The weights w_k stand for the prior probabilities of the components in the mixture and are derived from the results of the clustering stage, by computing the proportion of training images assigned to each cluster. In Gaussian formulation, the likelihood of the global features given the component is defined as:

$$p(\mathbf{g}|z_k) = N(\mathbf{g}; \mu_k^z, \Sigma_k^z) \tag{3}$$

with mean vector μ_k^z and covariance matrix Σ_k^z. Both parameters are obtained from the results of the clustering stage, by computing the parameters of the

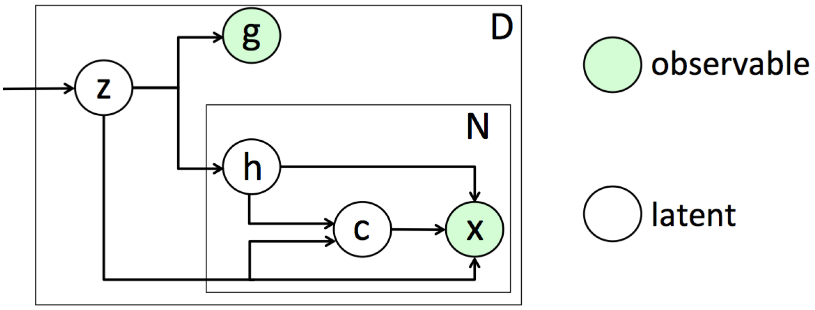


Figure 6: Graphical model of our approach Top-down visual attention modelling with manipulated objects. Nodes represent random variables, edges show dependencies among variables, and boxes refer to different instances of the same variable. Latent variables (transparent background): z set of global arms configurations, h set or arm labels, c set of hand centre positions. Observable variables (shaded-green background): g global features, x spatial locations. D and N refer respectively to the number of training images and number of pixels in a frame.

Gaussian distribution over the set of samples assigned to each cluster (global configuration).

After introducing our simplified model for global features, let us extend it by considering the distributions that depend on local features. For each elementary arms model z_k, we introduce the pdf of each hand $p(h|z_k)$, where h is an index variable with two possible values $h = 0, 1$ for left and right hand respectively. Once given the arms model z_k and the selected hand h, its local centre coordinates are also probabilistically modelled by the distribution $p(\mathbf{c}|h, z_k)$.

Finally, the likelihood of a point $\mathbf{x}$ belonging to the area of interest is expressed by the conditional distribution $p(\mathbf{x}|h, \mathbf{c}, z_k)$. This distribution models the probability of a pixel to belong to the object being manipulated given the current geometric configuration of arms and hands. It is easy to note that the relative object location and pose is different for various global configurations such as the ones shown in Fig. 3.

Putting everything together, we can define the partial model involving the *local features*:

$$p(\mathbf{x}|z_k) = p(h|z_k)p(\mathbf{c}|h, z_k)p(\mathbf{x}|h, \mathbf{c}, z_k) \tag{4}$$

Next, we can define the selected distributions for the local variables as:

1. The pdf $p(h|z_k)$ is given by an experimental discrete distribution $(p(h = j|z_k), j = 0, 1)$
2. The hand centre $\mathbf{c}$ follows a Gaussian distribution $p(\mathbf{c}|h = j, z_k) = N(\mathbf{c}; \mu_j^c, \Sigma_j^c)$.
3. The experimental pdf $p(\mathbf{x}|h = j, \mathbf{c}, z_k)$ is computed also on training set by superimposing all left and right hands from object segmentation images belonging to the cluster z_k.

The first two pdfs are simply learned by computing their parameters using samples on the training set (see sec. 2.2 for the details). For the third distribution $p(\mathbf{x}|h = j, \mathbf{c}, z_k)$, it becomes necessary to firstly crop object segmentation images (or Bounding Boxes instead of segmentation) by selecting a region around the hand centre, and then superimpose and accumulate all cropped object images belonging to the same global component. The resulting accumulated map for each hand and global configuration is then normalized to sum to one over spatial locations (to become a pdf).

Once in test, in order to compute the saliency map of a particular video frame, the learned distribution is accordingly shifted to the hand centre in the frame. Figure 7 shows different examples of these distributions for left and right hands, and a given five global appearance models.

Finally, integrating the distributions of global and local features, the *salience value of a pixel* $\mathbf{x}$ is defined as its likelihood over the proposed model for saliency:

$$\begin{aligned} S(\mathbf{x}) = p(\mathbf{x}, \mathbf{g}) &= \sum_{k=1}^{K} w_k p(\mathbf{g}|z_k)p(\mathbf{x}|z_k) \\ &= \sum_{k=1}^{K} w_k p(\mathbf{g}|z_k) \sum_{j=0}^{1} p(h = j|z_k)p(\mathbf{c}|h = j, z_k)p(\mathbf{x}|h = j, \mathbf{c}, z_k) \end{aligned} \tag{5}$$

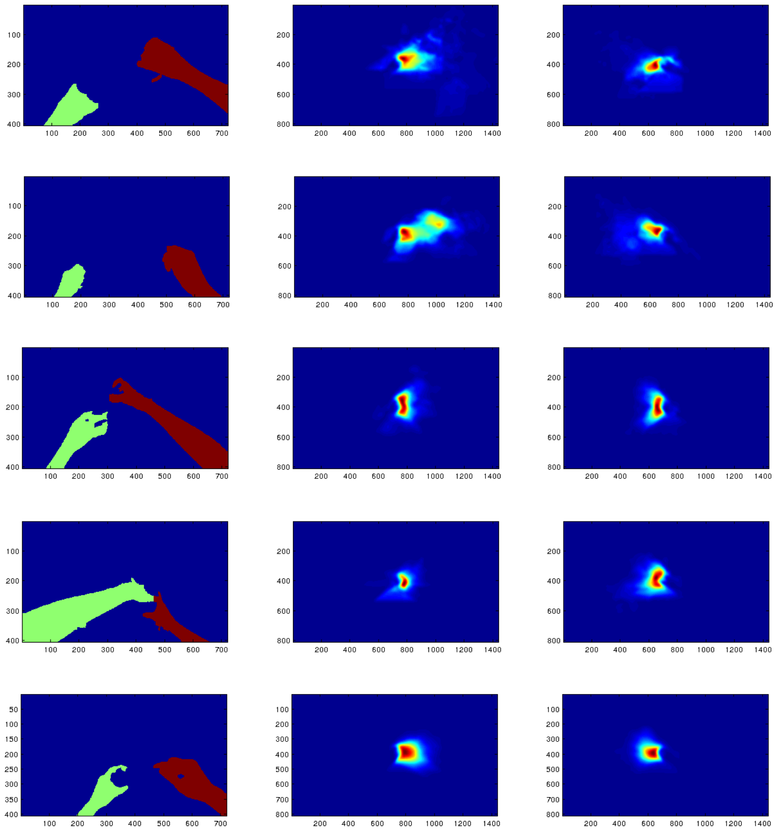


Figure 7: Five examples of the obtained experimental distributions $p(\mathbf{x}|h, \mathbf{c}, z_k)$. Left column: arm segmentation closest to cluster, Middle column: left hand distribution, Right column: right hand distribution.

Let us note that the model in eq.(5) allows to compute saliency even in the case where one of the arms is absent by simply considering the corresponding probabilities $p(h = 0|z_k)$ or $p(h = 1|z_k)$ as zero.

To summarize, we have developed a probabilistic model that explains how salient pixels are chosen based on hands/arms configuration and the relative expected location of the object being manipulated within each geometric arrangement.

3. Experiments and results

In this section we present the dataset and provide a whole description of the different experimental set-ups for the comparison of our probabilistic top-down saliency model against other saliency approaches. We also assess its contribution regarding manipulated object recognition performances.

3.1. Dataset description

The GTEA dataset we work on was introduced in [?]. 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 and objects of interest are provided for 17 videos. The frames were annotated with the objects of interest but we manually extended this annotation by drawing bounding boxes on them. The bounding boxes provide the "ground truth" results that could be reached with an "ideal" rectangular salient area. We did not use the setup proposed in [?], where the authors used videos from 3 subjects to train their system and the last one for evaluation, since the arm segmentations provided with the dataset do not cover all videos from Fathi's setup. Instead we have split the dataset into a training and test set of videos in such a manner as to even the number of samples of each object category in both sets.

For a better understanding, Table 2 contains the list of videos belonging to the training and test sets, Figure 8 shows the number of occurrences of each category in both sets. Let us note that this set-up can be consireded more challenging than the one presented in [?] since there is less training data and more test data. Furthermore we would also like to explain that, although videos from the same user are contained in both training and test datasets, it does not simplify the recognition task with respect to the original set-up as, in practice, both the scenario and manipulated objects are the same for every user in the dataset.

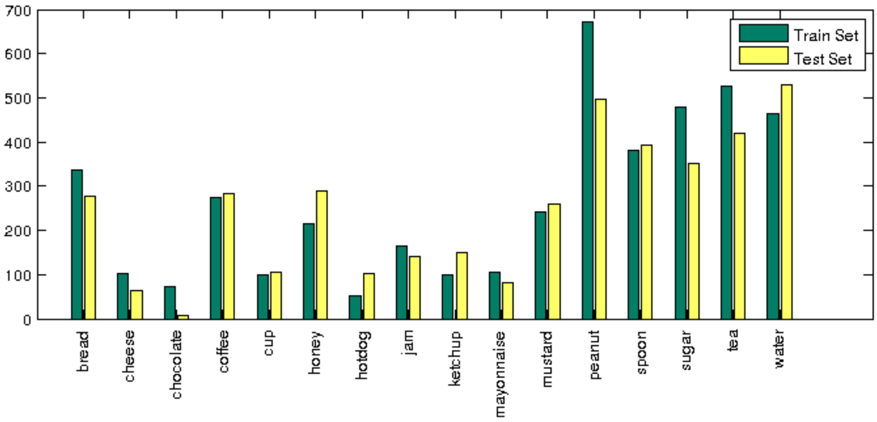


Figure 8: Occurences of each class in our Train and Test sets. The dataset has been split by videos so that the number of samples of each category in both sets is closest.

Training Set	**Test Set**
S3_Hotdog_C1	S3_Tea_C1
S1_Cheese_C1	S2_Tea_C1
S2_Peanut_C1	S2_Cheese_C1
S2_Coffee_C1	S2_Pealate_C1
S3_Coffee_C1	S1_Coffee_C1
S1_Tea_C1	S1_Hotdog_C1
S1_Pealate_C1	S1_CofHoney_C1
S3_Peanut_C1	S1_Peanut_C1
	S2_Hotdog_C1

Table 2: List of videos in Training and Test sets.

3.2. Selected visual saliency models for comparison

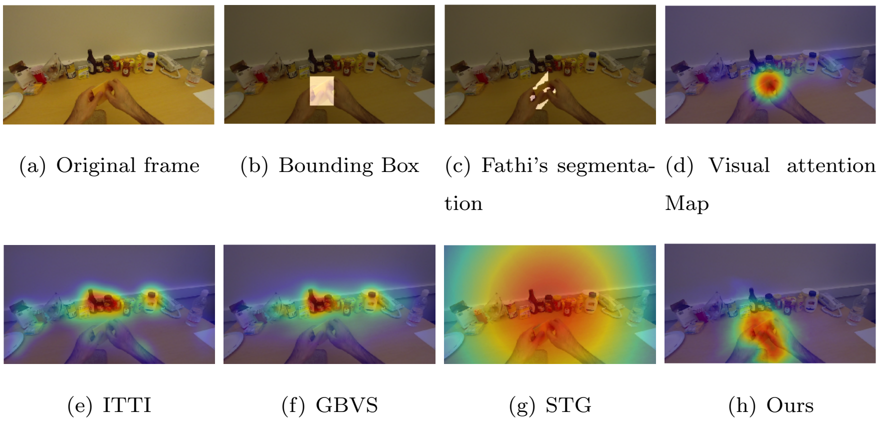
The following saliency prediction models were selected for comparison due to their popularity or particular suitability to egocentric video.

- The well-known reference model developed by Itti [**?**]. We will denote it as “ITTI” in the follow up of the paper.
- The graph-based visual saliency model developed by Harel [**?**]. It will now be referred to by the acronym “GBVS”.
- The spatio-temporal-geometric model presented in [**?**] since it has been specifically developed for saliency extraction in egocentric videos and presents the state-of-the art in saliency-based object recognition in this content [**?**]. This model will be referred as “STG”.
- Visual Attention maps built on gaze fixations by reference Wooding’s method [**?**]: the fovea projection for each fixation is modelled with a Gaussian of two visual degrees spread and resulting multi-Gaussian surface is normalized.

Figure 9 contains computed saliency maps for a randomly selected frame (a). We also display the manually annotated bounding box of the manipulated object (b), as well as the automatically extracted segmentation mask (c).

3.3. Psycho-visual evaluation of proposed saliency model

In this section we assess the capacity of our top-down model to predict human visual attention in the task-guided psycho-visual experiment. The saliency models presented in section 3.2 were also assessed for the sake of comparison. The psycho-visual experiment was designed for recording gaze fixations of subjects who observed the egocentric video with the task of recognition of manipulated objects. For this experiment 31 participants have been gathered, 10 women and 21 men. They were given a written instruction to look specifically at the manipulated object in videos. Each video was watched by at least 15 subjects. The gaze positions have been recorded with a HS-VET 250Hz Cambridge Research



(a) Original frame (b) Bounding Box (c) Fathi's segmentation (d) Visual attention Map

(e) ITTI (f) GBVS (g) STG (h) Ours

Figure 9: Saliency models selected for comparison.

Systems Ltd eye-tracker. The experiment conditions and the experiment room were compliant with the recommendation ITU-R BT.500-11 [**?**]. Videos were displayed on a 23 inches LCD monitor with a native resolution of 960 × 540 pixels. To avoid image distortions, videos were not resized to screen resolution but instead a grey frame was inserted around the displayed video. In order to avoid the visual fatigue, the duration of observation was not longer than 15 minutes for each subject.

Automatically predicted saliency maps can be compared to human gaze fixations with the help of dedicated metrics. From [**?**] and previous work [**?**], we retained the Normalized Scan Path (NSS) as the most frequently used and suitable for the comparison of saliency maps with human eye fixations:

$$NSS(p) = \frac{SM(p) - \mu_{SM}}{\sigma_{SM}} \tag{6}$$

where p is the location of one fixation and SM is the saliency map with its mean μ_{SM} and standard deviation σ_{SM}. The final NSS score is given by the average of the $NSS(p)$ values for all N eye fixations.

We measured the similarity of recorded eye fixations from the experiment with automatically generated saliency maps from our top-down probabilistic model and the ones presented in section 3.2. In total 8244 frames were compared

	ITTI	GBVS	STG	OURS
mean NSS score	1.05 ± 0.7269	1.29 ± 0.6551	1.52 ± 0.2490	$\mathbf{2.28} \pm 1.2226$

Table 3: NSS mean scores (with standard deviations) between human fixation points and different saliency map models. Our model outperforms the others

for each saliency model and the final mean scores with standard deviations are presented in Table 3. As shown in the table, our proposed top-down probabilistic model corresponds better to real human eye fixations than the other state-of-the-art saliency models. Since the standard deviation are high, we computed the p-values to back up the hypothesis that the NSS mean using our top down approach is significantly higher than with the other attention prediction models. At the 5% significance level, the data do provide sufficient evidence to conclude that the mean NSS score using our top-down saliency is greater than the mean obtained using other saliency models. It is however important to underline that the GVBS and ITTI models are bottom-up and were not designed for a task of recognition of specific objects of interest.

3.4. Object recognition performances

The ultimate goal of developing a model of top-down visual saliency is in the task of manipulated object recognition. Hence, we first present the object recognition approach with saliency-based psycho-visual weighting of features. This approach, combined with the proposed saliency model, is then compared to other state of the art paradigms for object recognition. We also benchmark it with other saliency models presented in section 3.2.

3.4.1. Saliency-based object recognition approach

In this study we used the saliency-based object recognition method presented in [**?**]. The approach is based on the well-known Bag-of-Visual Words (BoVW) paradigm [**?** **?**]. It uses dense SURF descriptors [**?**] and the BoVW is built when weighting each quantized feature by underlying predicted saliency value.

Since the saliency depends on the segmentation of the arms it is possible to find cases where arms do appear on the image or are not detected by the segmentation algorithm (this has happened only in 720 cases, meaning around 4.1% of all the segmentations provided by Fathi in the dataset). In these cases our model obviously does not provide a saliency map and it is up to the user to decide which saliency model to use. The models in section 3.2 constitute valid alternatives among which the STG ([**?**]) stands out as it has been specifically developed for saliency extraction in egocentric videos. In this work however, in order to rely solemnly on our model during the computation of performances, no other saliency model was computed to replace cases where arms are not detected. Instead we chose to build non-weighted signatures as in the original BoVW framework.

For the computation of BoVWs, we use a dictionary size of 4000 visual words. Once each image is represented by its weighted histogram of visual words, an SVM classifier [**?**] is used with χ^2 kernel. Posterior probabilistic estimates for the occurrence of the object of class C in the frame t are finally obtained using Platt's approximation [**?**].

3.4.2. Influence of the number of clusters in the global appearance model

The number of clusters K introduced in section 2.1 is an open parameter in our model. We have performed an optimization of the target mean Average Precision (mAP) of object recognition in regard to this parameter using the paradigm previously introduced in section 3.4.1. Table 4 below illustrates the influence of the number of clusters K to the target mAP. Having too few clusters might lead to a lack of information about certain arm models while having too many leads to poorly populated clusters. The case of $K = 1$ is the specific case where we do not consider the information given by global features. We observed that its high generality makes it perform well in most categories of objects. However, some categories of objects with specific shapes ("water", "ketchup", "sugar", . . .) are manipulated in certain ways such that removing global features yields a drop in recognition performances.

	$K = 1$	$K = 20$	$K = 50$	$K = 100$
mAP	0.301	0.316	**0.353**	0.342

Table 4: Validation of the number of global appearance models K

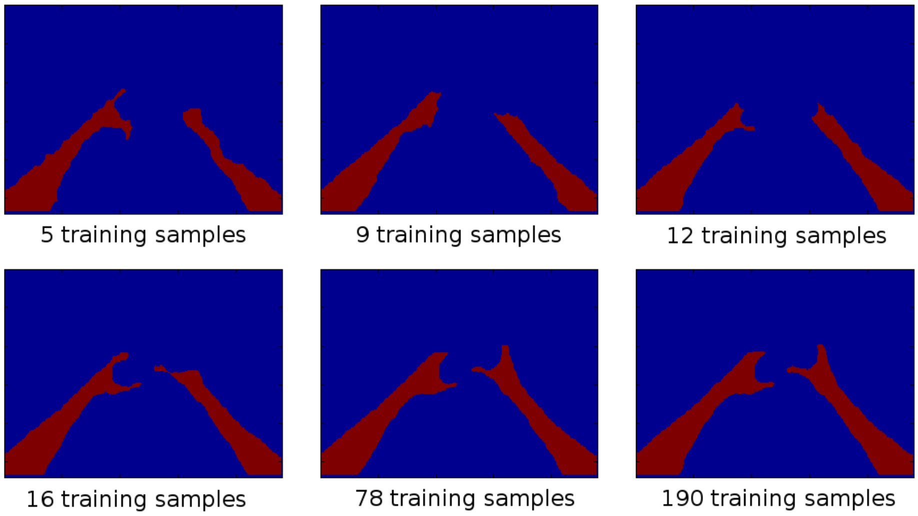
For the rest of the experiments the saliency model referred as "Ours" corresponds to the methodology presented in section 2 with $K = 50$ clusters, which has turned out to be the optimal value in our experiments.

3.4.3. Influence of the arm segmentation performances

As stated previously, this paper aims to provide a model for computing top-down semantic saliency maps given the arm segmentations. Hence the performance of our approach is deeply linked to the quality of the arm segmentation. In this part we aim to study how much segmentation errors could alter the performance of our proposed model. It is possible, based on the data provided in this dataset, to alter the given arm segmentations by applying varyingly important transformations to the previously segmented arms (e.g. homographies). However in order to truly degrade segmentation performances, we chose to implement a genuine hands segmentation framework and train it with different amount of training data.

Detection of hands/arms in egocentric videos has already been the core of several recent studies ([**?** **?** **?** **?**]). In this work, we retained the framework of [**?**] which has shown to provide good performances in similar contents. It is based on training modls for hand (arm) pixels with a training set of patches. Then a binary classification of pixel is performed. This segmentation paradigm was pioneer in the domain since it was the first to propose a model adapting to different illumination conditions, which proved to be essential in egocentric videos where lighting conditions vary often. Figure 10 shows some examples of how the segmentation gets affected by varying the number of training data.

In order to measure the segmentation performances 327 images were randomly chosen among the whole dataset and were manually segmented. We



5 training samples

9 training samples

12 training samples

16 training samples

78 training samples

190 training samples

Figure 10: Illustration of arm segmentation outputs with Li's model ([**?**]) for different amount of training data

therefore compared the similarity between automatic segmentation model and manually annotated data using the Jaccard's similarity coefficient:

$$J(S_m, S_a) = \frac{|S_m \cap S_a|}{|S_m \cup S_a|} \tag{7}$$

where S_m and S_a respectively stand for manual and automatic segmentation. Figure 11 shows the average similarity between Li's segmentation ([**?**]) and the 327 manual segmentations based on the amount of training data. We can see that the segmentation similarity with the ground truth grows with the amount of training data until convergence. The rise of performance is more pronounced for small numbers of training samples, and good similarity scores are rapidly reached (65% for 16 training samples). For 78 training samples, Li's segmentation obtains a similarity score equal to Fathi's, which gets even slightly outperformed for higher numbers of training samples until stabilization around a score of 71.5%. The standard deviation is however almost twice as small as the one obtained with Fathi's segmentation. The rationale behind is that Fathi's segmentation does not always detect arms leading to a Jaccard's coefficient of

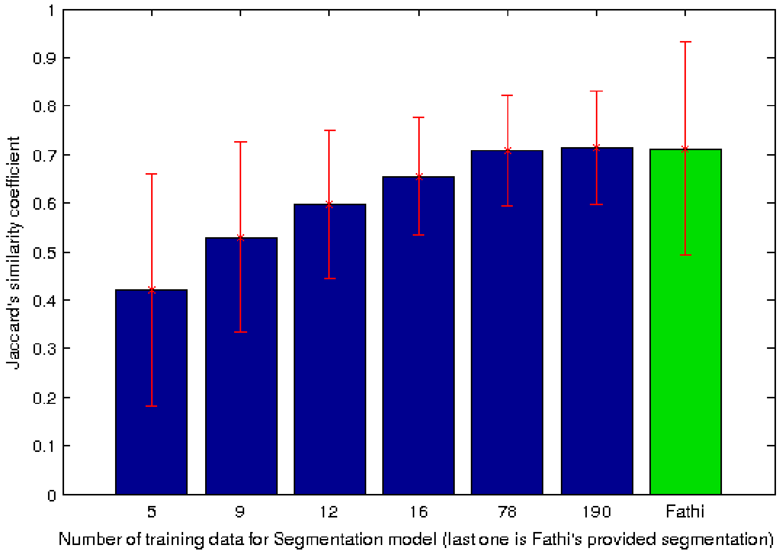


Figure 11: Average similarity between Li's segmentation ([?]) and the 327 manual segmentations based on the amount of training data. The last column is the similarity with Fathi's provided segmentation.

0 but, when it does, provides segmentations that are very close to the ground truth.

In Table 5 we present object recognition performances as mean Average Precision scores based on the number of data used to train the arm segmentation models. As expected, there is a significant drop of performance for low number of training data (and hence poor arm segmentation). A gap of more than 5% in mAP is noticeable between the lowest and highest object recognition scores. Two observations can be pointed out from these values however:

- As for the similarity scores in Figure 11, the variation of performance is not linear. We notice indeed that performances stay at their lowest point for segmentation similarity scores below 55% but abruptly raise and even start reaching convergence when getting closer to 60%.

- Even for a very low number of training data leading to notably poor arm

Number of training samples for Arm segmentation models	5	9	12	16	78	190
mAP	0.309	0.299	0.344	0.347	0.352	0.356

Table 5: Object recognition performances for different number of data used to train the arm segmentation models

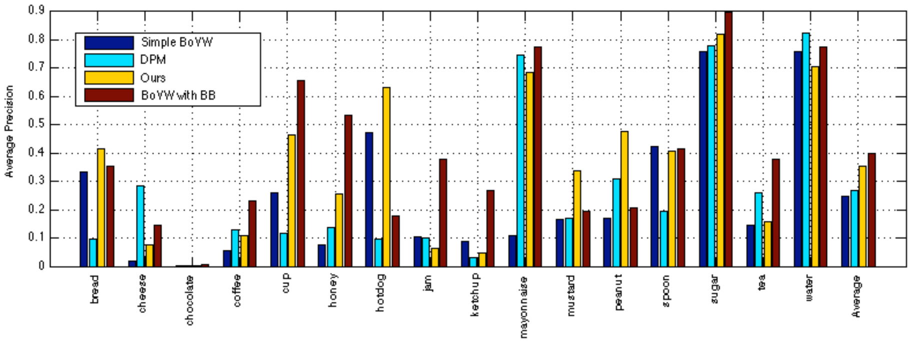


Figure 12: Object recognition performances between different paradigms. The results are given in average precision per category and averaged.

segmentation, our top-down model, coupled with the object recognition paradigm presented in section 3.4.1 still achieves higher performances with a simple BoVW framework (mAP of 0.246). This can be explained by the modularity of our model. Indeed, we observed that the K global Arm Models introduced in section 2.1 adapt to the poor segmentation by creating arm models even for these cases and learning an adequate experimental distribution $p(\mathbf{x}|h = j, \mathbf{c}, z_k)$.

3.4.4. Comparing with other object recognition approaches and saliency models

For the sake of comparison, we have compared our approach with a baseline model that implements a BoVW without any saliency maps, using a dense sampling of features on the whole frame. This method is referred as “Simple BoVW” in the experiments. In addition, we have also included in the comparison a “ground truth” model where descriptors were extracted only in manually annotated bounding boxes. In this method, referred as “BoVW with BB”, we consider the ground truth bounding boxes as “ideal” saliency maps.

Figure 12 shows the category detailed and average results for the object recognition. As can be seen from the mAP score (last set of bars), our method outperforms the two famous paradigms for object recognition in this kind of video content: i) it achieves an absolute improvements of 10.7% with respect to the base-line BoVW, and ii) a 8.6% absolute improvement with respect to the DPM model [**?**]. In addition, also achieves close performances to the “ideal” case, which was added for the upper bound estimate.

In section 2 we already raised the question of the need of building saliency maps if the objects have been already segmented. Indeed, segmentation as such cannot be used in our object recognition paradigm, as segmented objects are often represented by very small and sparse areas (see an example in Figure 9(c)). The extraction of relevant descriptors is thus strongly affected and yields a mAP of only 0.07. Nevertheless, one could object that segmented objects are too restrictive for a comparison to be possible. In this regard, we computed trivial saliency maps as a 2-dimensional fitted Gaussian on the segmented objects, allowing in the process to unify cluttered zones. Such saliency maps gave an object recognition mAP of 0.21, which is still very far from the score of 0.35 achieved by our model.

In their paper, Fathi et al. [**?**] use a different object recognition method based on the segmented zones. We also computed object recognition accuracy in our test set in the same way it was computed by Fathi. As can be seen in Figure 13 the precision obtained by our approach was slightly higher in average that Fathi’s. However it is important to note that the comparison is unfair since, as we already mentioned in section 3.1, we have evaluated our detectors under a more challenging set-up with less training data and more test data. Also an interesting thing to point out is how different both approaches detect better certain categories than other.

We also compare our model with those described in section 3.2 using the

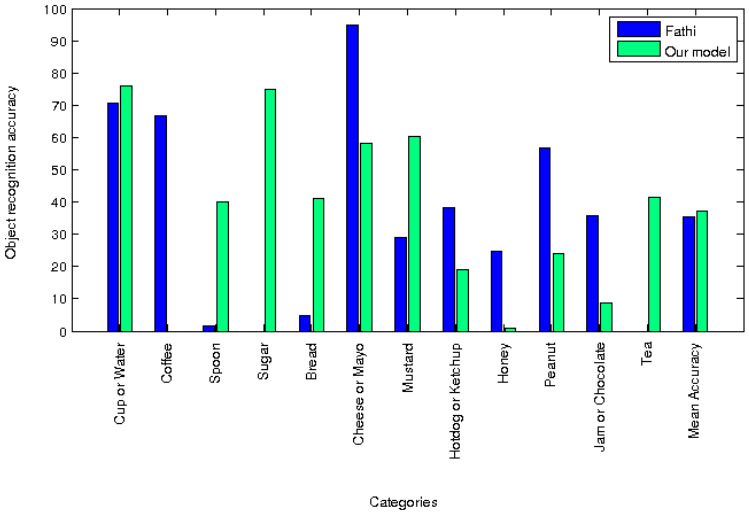


Figure 13: Object recognition accuracy comparison between the model presented in [**?**] and our approach.

same object recognition approach. Results for per-category and averaged object recognition are displayed in Figure 14 in terms of AP. Compared to ITTI and GBVS models, our model performs better for almost all categories. These bottom-up saliency models are stimuli-driven, make use of spatial contrast and were not designed to model a top-down, intentional attention component. The performances of bottom-up STG saliency maps, developed for video were also beaten for almost all categories. This is due to the overestimation by STG of the spread of Gaussian expressing central bias hypothesis on visual attention.

It also achieves slightly better performances than the ones provided by Human Visual Attention maps [**?**]. It is indeed better for some categories since as illustrated in Figure 9(d), the visual attention maps are perfectly located but sometimes do not cover the objects of interest enough, contrarily to our model (see Figure 9(h) for an example).

On Figure 14 we can see it is not necessarily that our top-down model outperforms other saliency methods in each category. We want to find out if the mean value of the population consisting in category APs (mAP) from

	ITTI/Ours	GBVS/Ours	STG/Ours
p-value	0.0876	0.0118	0.0541

Table 6: p-values between the population consisting in category APs (mAP) from our method and the ones obtained with APs of each of the other automated Saliency prediction methods (ITTI, GBVS, STG)

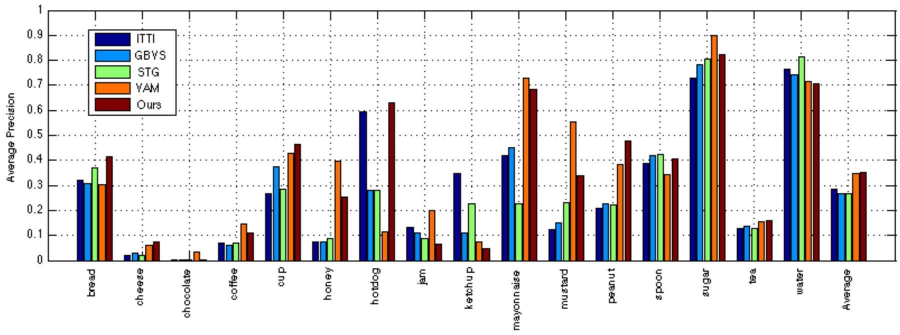


Figure 14: Object recognition performances between different saliency models applied to the saliency weighted BoVW paradigm.The results are given in AP per category and averaged.

our method is significantly different from the mAP obtained with each of the other automated Saliency prediction methods (ITTI, GBVS, STG). Hence we performed Student's t-tests with significance level of 0.10 for comparison and found the null hypothesis to be consistently rejected (p-values provided in table tab:pValuesSaliencies).

4. Conclusions and perspectives

In this paper we have proposed a top-down probabilistic visual saliency model for the target task of recognition of manipulated objects in egocentric video. It is based on global and local features and uses domain knowledge, i.e. the fact that the object of interest is manipulated by hands. The model predicts well human attention in a task-driven psycho-visual experiment and shows better performances than several bottom-up models widely used in literature, both in terms of comparison with human gaze fixations and target performance in manipulated object recognition task.

Despite the fact that this model has been developed for the specific case of egocentric video content and the task of manipulated object recognition, the idea behind is generic. It is indeed our belief that this model could be extended to other domains of application and not only egocentric videos with detection of arms. The model could be adapted to many scenarios where there exist reference objects, which can be easily recognized, and where the top-down attention is related to them. One interesting example is the aided robotic surgery or the generation of post-surgery video reports. Here, the reference objects are the medical instruments, so that the attention is driven to the close operation field. Further examples are the recognition of e.g. robot-sorted objects on a conveyor belt, carried objects by a crane in a surveillance scenario. Another example, again with egocentric video content, is the real-time detection of objects with wearable glasses for manipulation by neuro-prostheses. Anyway this is a general principle: task-driven visual attention can be easily predicted if we can detect the presence of reference objects for such a task, which in this paper were the hands of the user performing the action.

In visual attention modelling we need to use domain knowledge and contextual information. Visual attention is a complex combination of bottom-up, stimuli driven, and top-down, intentional components. In the perspective of the present research, combining of bottom-up and top-down prediction and spatio-temporal evolution of visual saliency in a video scene is envisaged with a target application to object and action recognition.

5. Acknowledgements

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

References

[1] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, in: IEEE Conference on Computer Vision and Pat-

tern Recognition (CVPR), 2012, IEEE, 2012.

[2] I. González Díaz, V. Buso, J. Benois-Pineau, G. Bourmaud, 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, MIIRH '13, ACM, 2013, pp. 11–14.

[3] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analyisis and Machine Intelligence 32 (9) (2010) 1627–1645.

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

[5] A. Fathi, Y. Li, J. M. Rehg, Learning to recognize daily actions using gaze, in: Proceedings of the 12th European conference on Computer Vision - Volume Part I, ECCV'12, Springer-Verlag, 2012, pp. 314–327.

[6] K. Ogaki, K. M. Kitani, Y. Sugano, 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, IEEE, 2012, pp. 1–7.

[7] C. Li, K. Kitani, Pixel-level hand detection in ego-centric videos, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013, pp. 3570–3577. `doi:10.1109/CVPR.2013.458`.

[8] Y. Pinto, A. R. van der Leij, I. G. Sligte, V. A. F. Lamme, H. S. Scholte, Bottom-up and top-down attention are independent, Journal of Vision 13 (3). `arXiv:http://www.journalofvision.org/content/13/3/`

16.full.pdf+html, doi:10.1167/13.3.16.
URL http://www.journalofvision.org/content/13/3/16.abstract

[9] M. Carrasco, Visual attention: The past 25 years,, Vision Research 51 (13) (2011) 1484–1525. doi:10.1016/j.visres.2011.04.012.

[10] L. Melloni, S. V. Leeuwen, A. Alink, N. G. Muumlller, Interaction between bottom-up saliency and top-down control: How saliency maps are created in the human brain., Cereb Cortex.

[11] A. Borj, L. Itti, State-of-the-art in visual attention modeling, IEEE Trans. on Pattern Analysis and Machine INtelligence 35 (1) (2013) 185–207.

[12] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[13] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 19, MIT Press, 2007, pp. 545–552.

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

[15] E. Vig, M. Dorr, D. Cox, Space-variant descriptor sampling for action recognition based on saliency and eye movements, in: A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid (Eds.), Computer Vision – ECCV 2012, Vol. 7578 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 84–97.

[16] A. T. T. Judd, F. Durand, A benchmark of computational models of saliency to predict human fixations, in: Computer Science and Artificial Intelligence Laboratory Technical Report, 2012.

[17] Y. F. Ma, X. S. Hua, L. Lu, H. Zhang, A generic framework of user attention model and its application in video summarization., IEEE Transactions on Multimedia 7 (5) (2005) 907–919.

[18] M. Cerf, J. Harel, W. Einhäuser, C. Koch, Predicting human gaze using low-level saliency combined with face detection., in: J. C. Platt, D. Koller, Y. Singer, S. T. Roweis (Eds.), NIPS, Curran Associates, Inc., 2007.

[19] T. Huawei, F. Yuming, Z. Yao, L. Weisi, N. Rongrong, Z. Zhenfeng, Salient region detection by fusion bottom-up and top-down features extracted from a single image, IEEE Transcations on Image processing 23 (10) (2014) 4389–4398. doi:10.1109/TIP.2014.2350914.

[20] D. Gao, S. Han, N. Vasconcelos, Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition., IEEE Trans. Pattern Anal. Mach. Intell. 31 (6) (2009) 989–1005.

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

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

[23] L. Itti, C. Koch, Computational modelling of visual attention, Nature Reviews Neuroscience 2 (3) (2001) 194–203.

[24] J. Li, Y. Tian, T. Huang, W. Gao, Probabilistic multi-task learning for visual saliency estimation in video, Int. J. Comput. Vision 90 (2) (2010) 150–165.

[25] D. Erhan, C. Szegedy, A. Toshev, D. Anguelov, Scalable object detection using deep neural networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[26] C. Shen, Q. Zhao, Learning to predict eye fixations for semantic contents using multi-layer sparse network, Neurocomputing 138 (2014) 61–68.

[27] M. Jones, , M. J. Jones, J. M. Rehg, Statistical color models with application to skin detection, in: Computer Vision and Pattern Recognition, CVPR'1999, IEEE Computer Society, 1999, pp. 274–280.

[28] L. S. Al, 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, DEPCOS-RELCOMEX '06, IEEE Computer Society, 2006, pp. 207–214.

[29] M. A. Fischler, R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395. `doi:10.1145/358669.358692`. URL `http://doi.acm.org/10.1145/358669.358692`

[30] H. Boujut, J. Benois-Pineau, R. Megret, Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion, in: A. Fusiello, V. Murino, R. Cucchiara (Eds.), Computer Vision – ECCV 2012. Workshops and Demonstrations, Vol. 7585 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 436–445.

[31] D. Wooding, Eye movements of large populations: Ii. deriving regions of interest, coverage, and similarity using fixation maps, Behavior Research Methods 34 (2002) 518–528, 10.3758/BF03195481.

[32] International Telecommunication Union, Methodology for the subjective assessment of the quality of television pictures, Recommendation BT.500-11, International Telecommunication Union (2002).

[33] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, T. Dutoit, Saliency and human fixations: State-of-the-art and study of comparison metrics, in: The IEEE International Conference on Computer Vision (ICCV), 2013.

[34] O. L. Meur, T. Baccino, Methods for comparing scanpaths and saliency maps: strengths and weaknesses., Behav Res Methods.

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

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

[37] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Comput. Vis. Image Underst. 110 (3) (2008) 346–359.

[38] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.

[39] J. C. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in: Advances in Large Margin Classifiers, MIT Press, 1999, pp. 61–74.

[40] S. Lee, S. Bambach, D. J. Crandall, J. M. Franchak, C. Yu, This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video, 2014.

[41] C. Li, K. Kitani, Model recommendation with virtual probes for egocentric hand detection, in: Computer Vision (ICCV), 2013 IEEE International Conference on, 2013, pp. 2624–2631. doi:10.1109/ICCV.2013.326.

[42] A. Betancourt, M. M. Lopez, C. S. Regazzoni, M. Rauterberg, A sequential classifier for hand detection in the framework of egocentric vision, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2014.