Saliency-based object recognition in video

Iván González-Díaz, Hugo Boujut, Vincent Buso, Jenny Benois-Pineau, Jean-Philippe Domenger

Saliency-based object recognition in video

Iván González-Díaz
Laboratoire Bordelais de
Recherche en Informatique
(LaBRI)
Talence, France
igonzale@labri.fr

Hugo Boujut
LaBRI
Talence, France
hugo.boujut@labri.fr

Vincent Buso
LaBRI
Talence, France
vbuso@labri.fr

Jenny Benois-Pineau
LaBRI
Talence, France
jenny.benois@labri.fr

Jean-Philippe Domenger
LaBRI
Talence, France
domenger@labri.fr

ABSTRACT

In this paper we study the problem of object recognition in egocentric video recorded with cameras worn by persons. This task has gained much attention during the last years, since it has turned to be a main building block for action recognition systems in applications involving wearable cameras, such as tele-medicine or life-logging. Under these scenarios, an action can be effectively defined as a sequence of manipulated or observed objects, so that recognition becomes a relevant stage of the system. Furthermore, video summarization tasks on such content is also driven by appearance of semantic objects in camera field of view.

One of the particularities of first-person camera videos is that they usually present a strong differentiation between active (manipulated or observed by the user wearing the camera) and passive objects (associated to background). In addition, spatial, temporal and geometric cues can be found in the video content that may help to identify the active elements in the scene. These saliency features are related to the modelling of Human Visual System, but also to motor coordination of eye, hand and body movements. In this paper, we discuss the automatic generation of saliency maps in video, and introduce a method that extends the well-known Bag-of-Words (BoW) paradigm with saliency information. We have assessed our proposal in several egocentric video datasets, demonstrating that it not only improves the BoW reference, but also achieves state-of-the-art performance of e.g. part - based models, with noticeably lower computational times. The approach has tremendous perspectives for other User Generated mobile Content.

Keywords

Egocentric vision, Object Recognition, Visual Saliency

1. INTRODUCTION

Recently, egocentric video analysis has gained a lot of interest due to the emerging end-user applications that involve the use of wearable cameras. Wearable cameras represent a cheap and effective way to record users' activity for scenarios such as telemedicine or life-logging.

In particular, the context of this work is a project that tackles the diagnosis, assessment, maintenance and promotion of self independence of people with dementia, such as Alzheimer disease. This objective requires to understand how the disease affects patients' activities in their lives, and to provide an objective assessment of their capacity to conduct the IADL (Instrumental Activities of Daily Living). Examples of early approaches addressing the same problem can be found in [20, 33]. In such a scenario, identifying human activities becomes a key problem to be solved, since it represents the basis to generate video semantic indexes that allow medical staff to efficiently navigate along the video footage.

Traditionally, the detection of human activities has been addressed by analyzing human motion patterns. More precisely, various approaches have successfully made use of the motion patterns associated to spatio-temporal interest points (STIP) in the video [24, 36]. In addition, the study of ego-motion has also resulted in successful approaches for first-person camera videos analysis [21].

However, in the particular case of egocentric view, we claim that an action can be effectively defined as a sequence of manipulated or observed objects, usually known as 'active' objects or 'objects-of-interest'. This assumption generally holds for video showing many household activities and, in particular, for the intended IADL scenario.

In that sense, the literature already shows examples in which the outputs of object detectors become the features for later action recognition systems in egocentric video: in [30], a vector with frame-level object probabilities is used for action recognition. A similar idea is explored in [13], where the authors propose a generative probabilistic approach that concurrently models activities, actions and objects. Furthermore visual object recognition in video is an open problem whatever the nature of the content.

In contrast to the well-known sliding window approaches for object detection and recognition [16, 22], and due to the specific nature of the first-person view contents, we aim to drive the object recognition process using *visual saliency*. Under the particular scenario of egocentric video, there is usually a strong differentiation between active (manipulated or observed by the user wearing the camera) and passive objects (associated to background) and, therefore, spatial, temporal and geometric cues can be found in the video content that may help to identify the active elements in the scene.

Incorporation of visual saliency in video content understanding is a recent trend. The fundamental model by L. Itti and C. Koch [18] is the most frequently used. Nevertheless other models can be proposed using priors on the content. The application of saliency modeling for object recognition on video serves for identifying areas where objects of interest are located. Then, features in these areas can be extracted for object description. Several works in the literature have shown the utility of human gaze tracking in the analysis of egocentric video content and, in particular, in the activity recognition task [14, 28].

This paper proposes an object recognition system that relies on visual saliency-maps to provide more precise object representations, that are robust against background clutter and, therefore, improve the precision of the object recognition task. We further propose to incorporate the saliency maps into the well known Bag-of-Words (BoW) [7] paradigm for object recognition. The benefits of this approach are multiple: a) the computation of saliency maps is generic (category-independent) and therefore a common step for any object detector, b) compared to sliding window approaches [8, 16], by looking at the salient area we can avoid much of the computationally overhead due to the scanning process and therefore use more complex non-linear classifiers, c) since the saliency maps are automatically computed in both training and test data, our method does not need bounding boxes for training, what dramatically reduces the human resources devoted to the database annotation.

We consider two differentiated scenarios of application. The first one is a *constrained* scenario in which all the subjects perform actions in the same room and, therefore, interact with the same objects: e.g. a hospital scenario in which the medical staff ask patients to perform several activities. This task can be seen as a specific object recognition problem since there is not intra-class variation between instances of a category other than this caused by the strong egomotion, changes on the viewpoint, illumination, occlusions,. . .

The second scenario, on the contrary, is *unconstrained*, and corresponds to recordings made at different locations. In this case users interact with various instances of the same objects: e.g. in a home environment, a patient performs daily activities using his/her own utensils and devices, that probably differ from those ones available in another home. The second scenario is therefore much more difficult than the first one, due to the large intra-class variation as well as to the limited amount of training data (a few instances of each object category).

In this paper, we will assess our method in both scenarios, showing its strength and weakness in comparison to other methods in the literature.

The remainder of the paper is organized as follows: in section 2 we discuss the related work. Next, in section 3, we provide a description of the geometric-spatio-temporal methods to compute visual saliency, and present some specially tailored developments to extend their use to an object recognition task in egocentric videos. Section 4 introduces our saliency-based approach for object recognition in egocentric video. In section 5 an in-depth evaluation is provided that assesses our model under the considered scenarios, and compares it to other state-of-the-art approaches. Finally, section 6 draws our main conclusions and introduces our further lines of research.

2. RELATED WORK

Object recognition is a very active task for the computer vision community. Initiatives such as the PASCAL Visual Object Classes (VOC) Challenge [11] promote the development of new algorithms by establishing a benchmark for model assessment in tasks like object recognition, detection and segmentation.

However, the particular problem of object recognition in egocentric video still lacks of enough specific approaches that exploit all the particularities of this type of content. Just to mention some of them: active objects in egocentric videos tend to appear in specific areas of the image, hands are the main source of occlusion, strong egomotion and object manipulation leading to dramatic changes on the viewpoint, active objects tend to appear at an approximately constant scale, . . .

Now looking at the specific field of object recognition in egocentric video, two kinds of approaches can be identified in the literature: those ones that rely on sliding windows, and those ones that first try to segment the foreground area to restrict the detection process.

Concerning the first type, the authors in [30] proposed to extend the use of the well known Discriminatively Trained Deformable Part-Based (DPM) Models [16], which has been demonstrated great performance in Pascal VOC [10], to egocentric video. As a part of their study, they have shown how object classifiers trained in general web-based collections such as ImageNet achieved very poor performance when applied to egocentric databases. The rationale behind is that objects appearance is different depending on the type of camera capturing the scene (e.g. 1st vs 3rd person camera view).

In what concerns the second kind of approach, the authors in [31] demonstrated how a figure-ground segmentation method helps to improve object classification. In particular, they used foreground segmentation masks to drive the object recognition process and evaluated their application to two object recognition systems: one using the sliding-window DPM [16], and another with exemplar-based SIFT matching [25].

A similar approach is followed in [15], which might be considered as the most advanced work towards the object recognition in egocentric views. In this paper, the authors 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 to the hands and the object being manipulated, respectively, and finally assigns an object label to the frame. This approach achieves impressive results when an object is manipulated and, furthermore, provides segmentation masks for each element. In contrast, to achieve their results, it requires active objects to be manipulated (to show other motion than the egomotion) and relies on a very complex process for foreground areas detection.

However, both methods using foreground segmentation show two main limitations: on the one hand, both of them assume that an active object moves arbitrarily in contrast to the background, that remains static in the world coordinate frame. From our point of view, this assumption is too constraining since many objects considered as 'active' in a scene can be also static. Some examples can be found in daily activities: a subject might be reading a book or a manuscript that is laid on a table and therefore not moving, it also might be watching TV (if we do not consider the residual motion inside the TV screen). Furthermore, even in cases when a user is manipulating an object, the object might look still for a notable segment of time. Next, both proposals consider binary segmentation masks to drive the recognition process so that the regions in the image are sharply considered as either relevant or irrelevant for the scene understanding.

Our method, in contrast, aids the recognition process using a soft measure based on visual saliency. Visual saliency has been successfully applied to object recognition in still images, either working on individual pixels as in [35], or computing saliency measures of bounding boxes, as in [1]. In our approach to object recognition in egocentric content, we follow temporal, but also spatial and geometric principles of visual saliency, making our method not restricted to the cases in which an object is manipulated by a user. Furthermore, rather than simply providing hard binary masks that delimit the area of a frame in which the recognition process is performed, we propose to use visual saliency 'soft' maps; hence, we do not simply filter out some areas of the image that are considered not salient, but weight the influence of each pixel in the recognition process. We believe that this approach successfully guides the recognition process to the areas of interest while still keeping the contribution of the context information around the object of interest, and in parallel, is more robust against errors in the saliency map estimation.

3. VISUAL SALIENCY FOR OBJECT RECOGNITION IN EGOCENTRIC VIDEOS

3.1 Spatial, geometric and Temporal Saliency Approaches

In order to drive the video analysis to the regions that are potentially interesting to human observers we need to model visual saliency on the basis of video signal features. In this work, we have considered three basic approaches to generate saliency maps, each of them built using a particular source of information: spatial, geometric and temporal. In the following paragraphs, we will briefly describe the method that gives place to each map.

Spatial saliency S_s: proposed in [5], it is based on various color contrast descriptors that are computed on the HSV color space, due to its closeness to human perception of color. In particular, 7 local contrasts are computed, namely:

1. *Contrast of Saturation*: A contrast occurs when low and highly saturated color regions are close.
2. *Contrast of Intensity*: A contrast is visible when dark and bright colors co-exist.
3. *Contrast of Hue*: A hue angle difference on the color wheel may generate a contrast.
4. *Contrast of Opponents*: Colors located at the hue wheel opposite sides create very high contrast.
5. *Contrast of Warm and Cold Colors*: Warm colors – red, orange and yellow – are visually attractive.
6. *Dominance of Warm Colors*: Warm colors are always visually attractive even if no contrast are present in the surrounding.
7. *Dominance of Brightness and Saturation*: Highly bright and saturated regions have more chances of attracting the attention, regardless of the hue value.

The spatial saliency value $S_s(i)$ for each pixel i in a frame is computed by averaging the outputs associated to the 7 color contrasts.

Temporal saliency S_t: this saliency models the attraction of attention to motion singularities in a scene. The visual attention is not grabbed by the motion itself, but by the residual motion for each pixel, e.g. the difference between the estimated motion for each pixel and the predicted camera motion based on a global parametrization.

Omitting many details, the process of computing a temporal saliency map is as follows: first, for each frame in the video, a dense motion map $\mathbf{v}(i)$ that contains the motion vectors in each pixel i in the image is computed using the optical flow technique described in [12].

Then, a 3x3 affine matrix A that models the global motion associated to the camera movements is computed. For that end, the well known robust estimation method RANSAC [17] has been used in order to successfully handle the presence of outliers (e.g. areas of the image associated to objects that move differently than the camera). Furthermore, since the central area of each frame constitutes the most likely region where moving objects appear, this region is not considered for the affine matrix estimation, thus reducing the proportion of outliers.

Next, the residual motion $\mathbf{r}(i)$ is computed by compensating the camera motion:

$$\mathbf{r}(i) = \mathbf{v}(i) - A\mathbf{x}_i \qquad (1)$$

where $\mathbf{x}_i$ stands for the spatial coordinates of each pixel i, $\mathbf{x}_i = (x_i, y_i, 1)^T$.

Finally, the values of the temporal saliency map $S_t(i)$ are computed by filtering the amount of residual motion in the frame. The authors of [5] reported that the human eye cannot follow objects with a velocity higher than $80^\circ/s$ [9]. According to this psycho-visual constraints, a post-processing filter was proposed in [5] that decreased the saliency when motion was too strong. Applying this filtering stage to our first-person camera videos was however too restrictive due to the strong camera motion so that we have preferred to consider a simpler filtering stage that normalizes and computes the saliency map as follows:

$$S_t(i) = \min\left(\frac{||\mathbf{r}(i)||_2}{K}, 1\right) \qquad (2)$$

where K has been heuristically computed depending on image dimensions (H,W), as $K = \max(H,W)/10$.

Geometric saliency S_g: it follows two observations about saliency in egocentric video: on the one hand, some studies on general purpose video confirm the so-called center bias hypothesis, that is the attraction of human gaze by the geometrical center of an image [4, 5]. On the other hand, in videos recorded with wearable cameras, the camera is usually set-up to point specific areas of interest: e.g. the gaze fixation if the camera is located in glasses, or an area just in front of the human body where the hands usually manipulate objects, in case it is located on the body. Generally, central geometric saliency is dependent on the wearable camera position and might be shifted in image plane [4]. In the present research, we work on datasets with either eye-centered or body-centered camera, thus using the center-bias hypothesis. Hence, following the approach in [5], the geometric saliency map $S_g(i) = \mathcal{N}((x_0, y_0), (\sigma_x, \sigma_y))$ is computed as 2D Gaussian located at the screen center with a spread $\sigma_x = \sigma_y = 5^\circ$.

However, this attraction may change with the camera motion. This is explained by the anticipation phenomenon [23]. Indeed, the observer of video content produced by a wearable video camera tries to anticipate the actions of the actor. The action anticipation is performed according to the actor body motion which is expressed by the camera motion. Hence we propose to simulate this phenomenon by moving the 2D Gaussian centered on initial *geometric saliency point* in the direction of the camera motion projected in the image plane. A rough approximation of this projection is the motion of image center computed with the global motion estimation model previously described.

Results on the basic approaches are shown in Figure 1 (columns 2-4). As one can notice from the figures, spatial and temporal saliency maps show more precise localization of the objects of interest whereas the geometric approach provides a coarse approximation of the visual saliency. However, saliency information appears more scattered or disaggregated for the first two approaches, being more compact and therefore robust for the geometric technique.

For an object recognition task, we consider that the perfect saliency map is a trade-off between precision and compactness, requirement that, based on the examples, is not completely fulfilled by any of the basic approaches. Hence, to overcome this issue, we

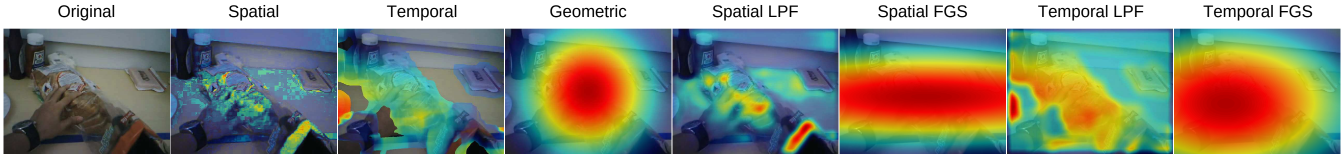


Figure 1: Results of various saliency maps for one frame in GTEA dataset. The three basic techniques spatial, temporal and geometric are shown. In addition, for spatial and temporal maps, two types of postprocessing are also included (LPF and FGS).

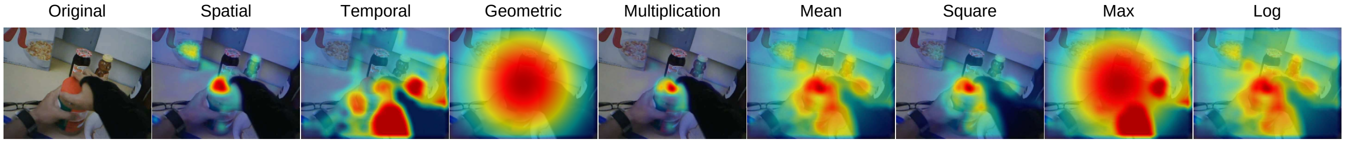


Figure 2: Results of various fusion strategies for computing spatio-temporal-geometric saliency maps.

propose two extensions: a) to incorporate a post-processing step on the spatial and temporal techniques that provides more compact saliency representations and, b) to investigate fusion schemes that successfully combine the three approaches taking advantage of their precision and compactness, respectively.

3.2 Postprocessing: Setting-up suitable saliency maps for object recognition

As already mentioned, we propose to use an additional post-processing stage to obtain more compact representations for the spatial and temporal saliency. In particular, we have evaluated two methods: a) a very simple spatial low-pass filtering using a Gaussian mask (LPF), and b) a method that fits a Gaussian Surface (FGS) on the original map.

The LPF approach, shown in columns 5-6 of Figure 1, simply provides a smooth version of the original saliency maps. However, if the standard deviation of the spatial Gaussian is large enough, results may fulfill our requirements of compactness.

For the second approach, given the original saliency mask S, we propose to fit a Gaussian surface of the form:

$$G(x,y) = Aexp\left[-\frac{1}{2}\left(\frac{x-x_g}{\sigma_x^2}+\frac{y-y_g}{\sigma_y^2}\right)\right] \qquad (3)$$

where $\theta = \{A, x_g, y_g, \sigma_x, \sigma_y\}$ are the parameters to be estimated in the fitting process. In practice, we minimize the square error between the two maps $e^2 = \sum_{x,y}[S(x,y) - G(x,y)]^2$ using the optimization method described in [27].

In the experimental section we will assess the performance of both post-processing approaches.

3.3 Fusion strategies for saliency maps

Once the basic spatial, temporal and geometric saliency maps has been introduced, we aim to evaluate how their combination into spatio-temporal-geometric saliency masks S_{stg} might improve the representation of the area of interest in the image.

For that end, several fusion strategies have been proposed and evaluated in this work. Again, although most of them have been already proposed in [3] in a video quality assessment task, for the sake of compactness we next briefly describe their computation:

1. Multiplication (Mult): a multiplicative fusion strategy model as:

$$S_{stg}^{mult}(i) = S_s(i) \times S_t(i) \times S_g(i) \qquad (4)$$

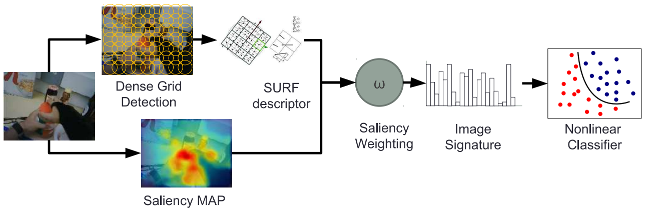


Figure 3: Processing pipeline for the saliency-based object recognition in first-person camera videos

2. Mean: the average of the three methods as:

$$S_{stg}^{mean}(i) = \frac{1}{3}(S_s(i) + S_t(i) + S_g(i)) \qquad (5)$$

3. Square: the squared Minkovsky pooling reinforced by multiplicative pooling:

$$S_{stg}^{sq}(i) = S_s(i) \times S_t(i) \times S_g(i) + \frac{1}{3}(S_s^2(i) + S_t^2(i) + S_g^2(i)) \qquad (6)$$

4. Max: maximum pooling:

$$S_{stg}^{max}(i) = \max(S_s(i), S_t(i), S_g(i)) \qquad (7)$$

5. Log: logarithmic combination model:

$$S_{stg}^{log}(i) = \frac{1}{3}(\log(1+S_s(i)) + \log(1+S_t(i)) + \log(1+S_g(i))) \qquad (8)$$

A visual example of the fusion strategies is shown in Figure 2. In addition, all of them will be evaluated in the experimental section of this paper.

4. A SALIENCY-BASED APPROACH FOR OBJECT-RECOGNITION

4.1 Low-level feature extraction and description

In this section we will describe our approach for object recognition in first-person camera videos using saliency masks. As we have already mentioned in the introduction, we aim to detect the region of interest (ROI) of each frame so that we can effectively build more precise image representations.

The processing pipeline of our approach is included in Figure 3. We build our model on the well-known Bag-of-Words (BoW) paradigm [7], and propose to add saliency masks as a way to improve the spatial precision of the original Bag-of-Words approach.

For each frame in a video sequence, we extract a set of N local descriptors using a dense grid of local circular patches [34]. Based on some experiments, we have set the radius of the circular patches to 25px, and the step size between each local patch of 6px, thus leading to a high degree of overlapping between neighboring local regions.

Next, each local patch $n = 1..N$ is described using a 64-dimensional SURF descriptor d_n [2], which has shown similar performance than the SIFT descriptor [25] in our experiments, whereas it is of half the dimension. Each descriptor d_n is then assigned to the most similar word $j = 1..V$ in a visual vocabulary by following a vector-quantization process. The visual vocabulary, computed using a k-means algorithm over a large set of descriptors in the training dataset (about 1M descriptors in our case), has a size of V visual words. As we will show in the evaluation section, we have experimented with visual vocabularies of different sizes V.

In parallel, our system generates a saliency map S of the frame with the same dimensions of the image and values in the range [0,1] (the higher the more salient is a pixel).

4.2 Object recognition with Saliency Weighting

In the traditional Bag-of-Visual-Words approach [7], the final image signature H is the statistical distribution of the image descriptors according to the codebook. This is made by first assigning each local descriptor to a visual word in the vocabulary and then computing a histogram of word occurrences by counting the times that a visual word appears in an image.

Instead of doing this hard assignment, we propose to apply what we call *saliency weighting*, a sort of soft-assignment based on saliency maps. With saliency weighting, the contribution of each image descriptor is defined by the maximum saliency value found under the circular region Ω_n associated to the index n. In other words, descriptors over salient areas will get more weight in the image signature than descriptors over non-salient areas. Therefore, the image signature is a V-dimensional vector H that can be computed as follows:

$$H_j = \sum_{n=1}^{N} \alpha_n w_{nj} \tag{9}$$

where the term $w_{nj} = 1$ if the descriptor or region n is quantized to the visual word j in the vocabulary, and the weight α_n is defined as:

$$\alpha_n = \max_{s \in \Omega_n}(S(s)) \tag{10}$$

where Ω_n represents the set of pixels contained in the n circular region of the dense grid, and $S(s)$ is a saliency map.

Finally, the histogram H is L1-normalized in order to produce the final image signature.

It is worth stressing the difference between our weighted histogram with hard-assignments and the histogram with soft assignments previously proposed in the literature [19]. In that work, given a descriptor, a similarity measure is computed with respect to all the words in the vocabulary so that various bins of the histogram can be incremented according to these similarities. On the contrary, our method is assigning each descriptor to just one word in the vocabulary but then is weighting its contribution to the histogram using the saliency map information. In fact, if necessary, our method might be combined with the one in [19].

On the contrary, our method of saliency weighting is more similar to the spatial weighting proposed in [26] but, in our case, the weights are computed unsupervisely, without need of training data and not depending on the category to detect.

Once each image is represented by its weighted histogram of visual words, we use a non-linear classifier to detect the presence of a category in the image. In particular, we have employed a SVM classifier [6] with a χ^2 kernel, which has shown good performance in visual reignition tasks working with normalized histograms as those ones used in the BoW paradigm [32].

5. EXPERIMENTAL SECTION

In this section we assess our model in various challenging datasets with egocentric videos. As we have already, mentioned we aim to recognize objects under two different scenarios: constrained, in which all videos contain the same instances of the involved object categories, and the unconstrained, in which each video shows a different environment with varying instances of the object categories.

5.1 Datasets

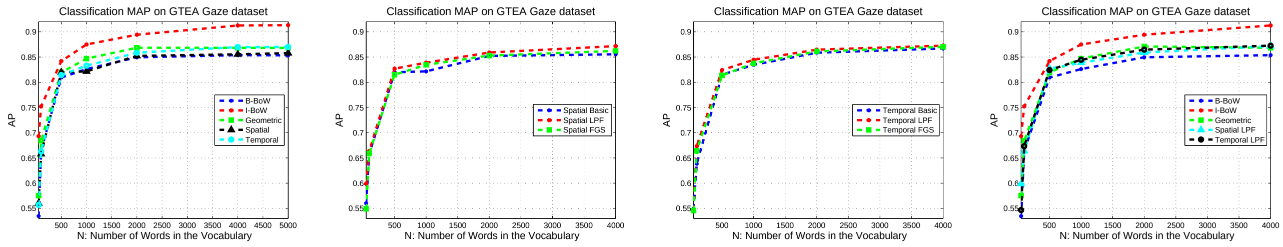
We have assessed our approach with three publicly available egocentric video datasets.

The first one is the GTEA Gaze dataset [14], which consists of 17 standard definition (640x480) video sequences, captured at a frame rate of 15 frames per second, and performed by 14 different subjects using Tobii eye-tracking glasses. Due to the lack of object annotations in this dataset, we have extracted and annotated 595 frames from the videos so that we can easily perform our tests over a set of still images. The whole dataset has been divided into two sets, namely: a) the training set (294 frames), and b) the test set (300 frames). Furthermore, we aimed to detect 15 object categories in this database. Due to its limited size, we have used this dataset to compare various system configurations.

The second dataset is the GTEA dataset [15] for Object Recognition. This dataset, recorded at 30 frames per second in 1280x720 definition, contains 7 types of daily activities, each performed by 4 different subjects. In this case, the camera is mounted on a cap worn by the subject. Weak annotations are already available for this dataset. They identify active objects on each frame belonging to 16 object categories, but do not include the object location. Since all the users have been recorded in the same room interacting with the same objects, we have evaluated our constrained scenario using this dataset. For that end, we have followed the same setup described in [15], using the users 2-4 for training the algorithms and the user 1 for testing.

The third dataset used in the experiments is the ADL dataset [30], that contains videos captured by a chest-mounted GoPro camera on users performing various daily activities at their homes. The high definition videos (1280x960) are captured at rate of 30 frames per second and with 170 degrees of viewing angle. In total, 27,064 frames have been accurately annotated providing bounding boxes for objects belonging to 44 categories. In our experiments, we have just considered those objects labelled as 'active' (those being interacted or observed by the users) for both training and testing purposes. This dataset is more challenging than the other two since both the environment and the object instances are completely different for each user, thus leading to an unconstrained scenario. However, we have evaluated both scenarios with this dataset: the constrained one by randomly dividing the whole set of frames into a training and test set (50/50%), and the unconstrained, by doing so at the video/user level.

5.2 Setting-up the final model



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

Figure 4: A comparison of various configurations in the GTEA Gaze dataset and various vocabulary sizes. (a) Results of the basic saliency techniques in comparison with the two references; (b) results achieved by two post-processing techniques for the spatial saliency; (c) results achieved by two post-processing techniques for the temporal saliency; (d) a comparison between the best post-processing option (LPF) and the reference methods.

In this section we compare various system configurations on the GTEA Gaze dataset. The objective is then to select the final system setup that provides the best performance, which will be compared with other state-of-the-art methods in the two envisaged scenarios.

5.2.1 Evaluating the basic approaches for saliency maps

We have firstly evaluated our basic approaches for generating the saliency maps. In addition, we have included two reference methods in the comparison:

1. Basic BoW (B-BoW): the Bag of Words approach that generates image signatures considering whole images. This method becomes the basic reference and allows us to evaluate the improvement achieved by our saliency masks.

2. BoW with Ideal Masks (I-BoW): this approach makes use of the ideal ground truth masks provided in the annotation. Since it evaluates our approach when the saliency masks correspond with the ground-truth, it constitutes the theoretical limit in its performance. It is worth noting how this ideal binary masks are used both on training and testing, thus incorporating the annotations in the whole recognition process, but omitting the aforementioned weighting scheme in the histograms computation.

The results of this study in the GTEA Gaze dataset are presented in Figure 4(a), that shows the Average Precision (AP) achieved by each approach at various vocabulary sizes. As one can notice from the results, for almost every technique, the performance improves until a vocabulary size of V=4000 words, after which it stabilizes. Hence, from now on, we will either remove larger vocabulary sizes from our experiments or simply consider the optimal vocabulary size of 4000 as the final approach.

Comparing the approaches, as we expected, the I-BoW constitutes the theoretical upper bound of the method. This is logic due to the use of the ground-truth bounding boxes that, although do not correspond to the tight silhouette of the object of interest, always ensure its correct localization. Furthermore, two of the basic techniques to compute the saliency masks (geometric and temporal) already achieve slightly better results than the reference B-BoW. This is a nice consequence of the use of saliency masks, even when not specific post-processing is applied to the the maps. Furthermore, the fact that the geometric saliency map is the one that achieves the best results, let us to conclude that compactness is even more important than localization precision for an object recognition task.

5.2.2 Techniques for saliency map post-processing

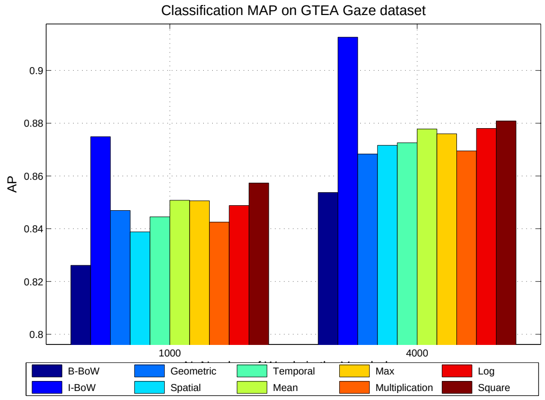


Figure 5: Classification results of various strategies for fusing spatio-temporal saliency maps. Values are given at two different vocabulary sizes (V=1000,V=4000). Basic and reference methods are also included for comparison.

In this section, we present the evaluation of the post-processing techniques described in section 3.2. As we have already claimed, direct outputs from some saliency detectors might not be optimal for an object recognition task due to the lack of compactness.

Since the Geometric technique already provided compact and Gaussian-shaped saliency masks, we have applied the postprocessing stage to the spatial and temporal techniques. Figures 4(b) and 4(c) respectively compare the results obtained in the GTEA Gaze dataset by the basic spatial and temporal saliency, and the two post-processing methods: Low Pass Filtering (LPF) and Fitting of a Gaussian Surface (FGS). The improvements on the results, although not very notable, demonstrate that post-processing is important to adequate the saliency maps to the particular problem of object recognition. Furthermore, the computational cost of the LPF method, the one that achieves the best performance, is almost negligible when compared to other steps of the processing pipeline.

In addition, Figure 4(d) shows a comparison between the LPF approach and the two reference methods. With the post-processing stage, now all the saliency methods outperform the reference B-BoW and achieve closer results to the theoretical limit I-BoW. Hence, from now on, LPF post-processing will be incorporated to every version of our approach.

5.2.3 Fusion techniques for saliency maps

Table 1: mAP and standard deviation on ADLdataset under the constrained and unconstrained scenarios.

Algorithm	Cons. mAP ± std	Uncons. mAP ± std
B-BoW	0.585 ± 0.258	0.113 ± 0.152
I-BoW	0.621 ± 0.250	0.191 ± 0.258
DPM [16]	0.341 ± 0.254	0.129 ± 0.194
Proposal	0.602 ± 0.260	0.125 ± 0.167

We have also evaluated the fusion approaches described in section 3.3.

Results of this experiment are shown in Figure 5. All the fusion strategies achieve better results than the basic approaches except for the multiplicative technique. The rationale behind is that this strategy is too restrictive and requires all basic saliency maps to show significant values in order to consider a pixel as salient.

The square fusion strategy obtains particularly good performance on this dataset, outperforming both the basic saliency approaches and the rest of the fusion strategies. In particular, by using this approach we are achieving absolute gains with respect to the reference B-BoW of a 3.1% and 2.7%, for a vocabulary of size 1000 and 4000, respectively. Hence, we will consider this fusion strategy as the final choice for our object recognition system in ego-centric videos.

5.3 Experiments under the constrained scenario

As we mentioned before, the constrained scenario is that one in which all the subjects wearing cameras are recorded in the same environment and interacting with the same object instances.

Results for the ADL dataset under the constrained scenario are shown in the first column of Table 2 in terms of mAP (mean Average Precision), and its standard deviation (category deviation). It is worth noting that we show only the results of those objects considered as 'active' in the dataset ground-truth annotations, e.g. those objects that are either manipulated or observed by the main actor in the ego-centric video. We consider these objects as the main source of information for detecting an action, so that the rest of the visual information (background) is less relevant and only useful for horizontal tasks as context identification.

As we have already mentioned, to simulate the constrained environment, we have randomly divided the whole set of frames into a training and test set (50/50%) without taking into account the video to which each frame belongs. In this dataset, we are comparing the performance of our approach with the reference method $B-BoW$, the ideal case $I-BoW$, and the Discriminatively Trained Part-Based Model (DPM) [16], which was the approach used by the authors of the dataset [30] to address the object recognition task.

Furthermore, in Figure 6 we include detailed per-category performance. Base on these results, we can draw the following conclusions:

- Our proposal outperforms the reference B-BoW by guiding the recognition process to the salient areas of each frame. This result is consistent along almost all the categories in the dataset, and supports the idea that using visual saliency generates more accurate object representations and reduces the effect of clutter.

- The approach using ideal masks is, as expected, the one yielding the best performance. However, a deeper by category analysis shows remarkable conclusions: in general, providing an accurate localization of the object (I-BoW) helps the recognition process and improves the performance. This observation is particularly noticeable for relatively small objects such as the ones belonging to the categories 'foodsnack', 'knife_spoon_fork', 'milk_juice' or TV. However, when the objects are too small, such as the instances of 'comb', 'dentfloss' or 'pills', we have observed that the ground truth bounding boxes, restricted to the object and lacking any information about object context, give not enough information to successfully detect its presence. In contrast, due to the fact that the saliency maps usually cover more area in the image (object, hands, even spatial neighboring context), our proposal achieves notably better results than the I-BoW. In addition, the reference B-BoW also achieves better results than I-BoW for these classes, although its performance is still below our approach.

- The performance of the DPM is poor when compared any BoW method. From our point of view, the rationale behind is that this method has been designed to get good generalizations of object categories, what prevents from taking advantage of the high visual similarity between training and test samples in the constrained scenario. Hence, we believe that its relative performance with respect to our approach should drastically improve in the unconstrained scenario.

In addition, we have also evaluated our approach in the GTEA dataset. This dataset represents the constrained scenario in a more realistic way, due to the fact that we can take training and test samples from different videos. Hence, we have followed the same evaluation setup proposed by the authors [15]. In particular, we have developed a multiclass classifier so that each image is considered to contain just one object of interest. Our proposal achieves a global classification accuracy of 36.8% in this dataset, which compares well with the 35% obtained by the authors of the dataset [15] when they matched the highest detection score to the ground truth annotations.

5.4 Experiments under the unconstrained scenario

The unconstrained scenario corresponds to the challenging situation in which users perform their activities at several locations, thus interacting with heterogeneous instances of the object categories. Consequently, the large intra-class variation jointly with the reduced number of object instances, are expected to lead to poor generalization in recognition process.

In our experiments, we have used the videos corresponding to half of the subjects $\{2, 3, 5, 7, 8, 12, 13, 14, 17, 18\}$ for training, and the remainder videos for test.

Average results of this study are shown the second column of Table 2, whereas Figure 7 shows detailed per-category AP. We next draw the main conclusions of this experiment:

- As expected due to the challenging nature of this scenario, the performance is drastically lower for all the automatic approaches (from AP ~ 0.6 to AP ~ 0.10). This illustrates how challenging is the problem of object recognition when just a few instance are available for each object.

- Furthermore, the I-BoW, that uses ground-truth masks in test, now notably outperforms any automatic approach. This fact stresses the importance of a good previous localization of the object of interest for its localization.

- Our proposal again outperforms the basic reference (B-BoW). The improvement is once more consistent along almost all the categories.

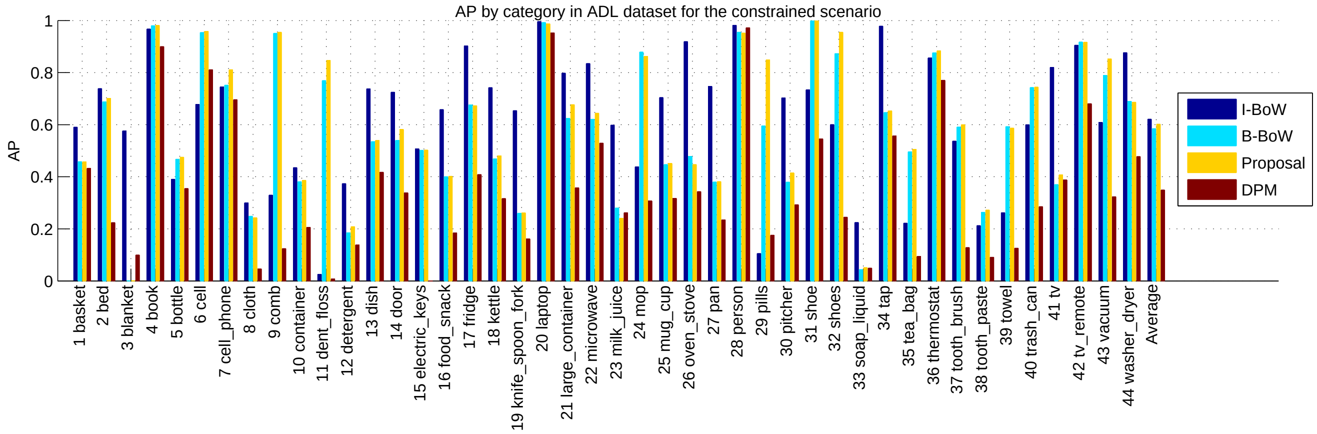


Figure 6: Per-category results (AP) for the constrained scenario achieved by various methods in the ADL dataset

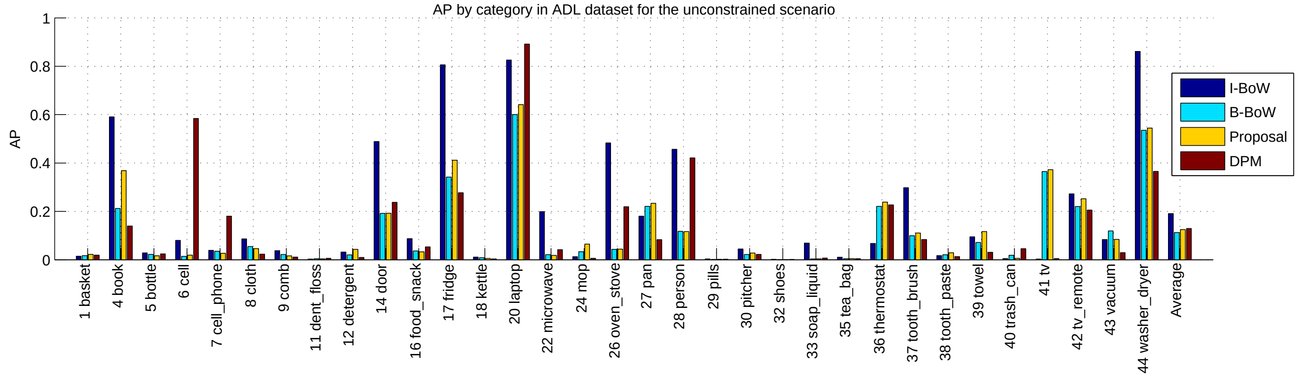


Figure 7: Per-category results (AP) for the unconstrained scenario achieved by various methods in the ADL dataset. Some categories cannot be computed in this scenarios due to the lack of samples in training/test sets.

Table 2: Test execution times of our approach compared with the DPM implementation in [16]. We show single threading (S.T.) and multi-threading (M.T.) execution time.

Algorithm	S.T.	M.T.
DPM [16]	60.4s	10.9s
Proposal	15.7s	4.1s

- The DPM now achieves competitive results, even slightly superior to the ones of our proposal. As we previously stated, this technique learns object models with a high degree of generalization, which is better suited for this unconstrained rather than for the constrained scenario.

As a conclusion, we can state that our approach yields good results in the constrained scenario, outperforming state-of-the-art approaches, and obtains competitive results in the unconstrained one. In addition to the classification results, it offers two main advantages over other alternatives: 1) it does not require precise localization of objects in the training data, what minimizes the human effort in the database annotation, and 2) as we will see in the next section, its computational complexity is low when compared to sliding window methods.

5.5 A study of the computational time

In Table 2, we show a comparison between the average execution times of our proposal and the DPM to run one category object-detector in a test frame. We included results using a single threading (S.T.) and multi-threading in a 2.10GHz computer with 4 cores, and hyper-threading enabled.

For our proposal, the execution time comprises the generation of the saliency maps, the SURF feature extraction process, the computation of the weighted histograms, and the classification using a SVM with χ^2 kernel. It is worth noting that some of the computations for the spatial saliency map are implemented in GPU so that they cannot be translated to S.T. case (spatial saliency takes about 0.05 sec per frame in the GPU). The rest of the calculus are made with the CPU under the aforementioned circumstances.

For the DPM, we run the implementation in [16], made in Matlab with optimized c routines for all the steps in the process that require most of the execution time.

As we can see in the tables, our approach shows much lower computational times in comparison with DPM. From our point of view, the rationale behind is the fact that using the saliency maps, we avoid the heavy scanning process of a sliding window approach as the DPM.

Furthermore, it is also worth noting that, since the saliency maps are automatically computed in both training and test data, our method does not need bounding boxes for training, what dramatically reduces the human resources devoted to the database annotation when compared to the DPM.

6. CONCLUSIONS AND FUTHER WORK

In this paper we have presented a method for object recognition in egocentric videos. Our proposal aims to drive the recognition process using visual saliency. In particular, spatial, temporal and geometric cues found in egocentric videos are exploited to improve the object recognition, generating more precise representations of the area of interest in a frame, as well as enhancing the robustness against cluttered backgrounds.

We have also evaluated several fusion strategies to generate spatio-temporal-geometric saliency maps from their basic constituents, as well as some post-processing techniques that improve the compactness, a property that has turned out to be very important for object recognition.

In addition, rather than simply performing foreground/background segmentation to restrict the recognition process to the areas of interest, we have proposed a soft application of saliency that controls the influence of pixels in the final object representation based on their saliency. We have combined saliency with the well known Bag of Words paradigm by proposing a saliency weighting method to compute image signatures.

Having in mind the context of this work, which is the automatic analysis of videos for the diagnosis, assessment, maintenance and promotion of self-independence of people with dementia, we have assessed our model in two particular scenarios of interest: a) a constrained scenario in all the subjects perform actions in the same room and, therefore, interact with the same object instances, and b) an unconstrained scenario that corresponds to recordings made at different locations, so that users interact with various instances of the same objects.

Our experiments have shown that this method outperforms the basic BoW model and achieves closer results to an hypothetical case in which optimal foreground masks are available in test. Furthermore, our approach compares well, and outperforms DPM and the full method in [15] under the constrained scenario. Furthermore, the computational time is less than half of the DPM one.

However, the notable decrease in performance in case of an unconstrained scenario reveals that our method needs further development. Indeed, in an unconstrained scenario the variability of object instances intra-category requires drastically new recognition approaches. Here we are in the case of "concept recognition". As we know from e.g. TRECVID challenge [29] concept recognition is a complex and open research problem and we are amongst those working on it.

7. REFERENCES

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 34(11):2189–2202, 2012.

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

[3] H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, and P. Bonnet. A metric for no-reference video quality assessment for hd tv delivery based on saliency maps. In *IEEE International Conference on Multimedia and Expo*, july 2011.

[4] H. Boujut, J. Benois-Pineau, and R. Megret. Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion. In A. Fusiello, V. Murino, and R. Cucchiara, editors, *Computer Vision âĂŞ ECCV 2012. Workshops and Demonstrations*, volume 7585 of *Lecture Notes in Computer Science*, pages 436–445. Springer Berlin Heidelberg, 2012.

[5] 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*, 2009.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[7] 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*, pages 1–22, 2004.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In C. Schmid, S. Soatto, and C. Tomasi, editors, *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005.

[9] S. J. Daly. Engineering observations from spatiovelocity and spatiotemporal visual models. In *IS&T/SPIE Conference on Human Vision and Electronic Imaging III*, 1 1998.

[10] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[12] G. Farnebäck. Fast and accurate motion estimation using orientation tensors and parametric motion models. In *Proceedings of 15th International Conference on Pattern Recognition*, volume 1, pages 135–139, Barcelona, Spain, September 2000. IAPR.

[13] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In D. N. Metaxas, L. Quan, A. Sanfeliu, and L. J. V. Gool, editors, *International Conference in Computer Vision, ICCV 2011*, pages 407–414. IEEE, 2011.

[14] A. Fathi, Y. Li, and 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, pages 314–327, Berlin, Heidelberg, 2012. Springer-Verlag.

[15] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 3281–3288. IEEE, 2011.

[16] 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*, 32(9):1627–1645, 2010.

[17] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24:381–395, June 1981.

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

[19] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in

large scale image databases. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 –8, June 2008.

[20] S. Karaman, J. Benois-Pineau, R. Mégret, V. Dovgalecs, J.-F. Dartigues, and Y. Gaëstel. Human daily activities indexing in videos from wearable cameras for monitoring of patients with dementia diseases. In *International Conference on Pattern Recognition (ICPR), 2010*, pages 4113 –4116, aug. 2010.

[21] K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3241 –3248, june 2011.

[22] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*. IEEE Computer Society, 2008.

[23] M. Land, N. Mennie, and J. Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28:1311–1328, 1999.

[24] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, pages 1 –8, june 2008.

[25] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[26] M. Marszałek and C. Schmid. Spatial weighting for bag-of-features. In *IEEE Conference on Computer Vision & Pattern Recognition*, volume 2, pages 2118–2125, jun 2006.

[27] J. J. Moré and D. C. Sorensen. Computing a trust region step. 4(3):553–572, Sept. 1983.

[28] 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*, pages 1–7. IEEE, 2012.

[29] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. QuÃl'enot. Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[30] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012.

[31] X. Ren and C. Gu. Figure-Ground Segmentation Improves Handled Object Recognition in Egocentric Video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[32] V. Sreekanth, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Generalized RBF feature maps for efficient detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2010.

[33] D. Szolgay, J. Benois-Pineau, R. Megret, Y. Gaestel, and J.-F. Dartigues. Detection of moving foreground objects in videos with strong camera motion. *Pattern Analysis and Applications*, 14:311–328, 2011.

[34] T. Tuytelaars, C. Lampert, M. Blaschko, and W. Buntine. Unsupervised object discovery: A comparison. *International Journal of Computer Vision*, 88:284–302, 2010.

[35] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *IEEE International Conference on Computer Vision*, 2009.

[36] S.-F. Wong and R. Cipolla. Extracting spatiotemporal interest points using global information. In *International Conference on Computer Vision, 2007. ICCV 2007.*, pages 1 –8, oct. 2007.