Visual search for objects in a complex visual context: what we wish to see

Hugo Boujut, Aurélie Bugeau, Jenny Benois-Pineau

1

Visual search for objects in a complex visual context: what we wish to see

Hugo Boujut

University of Bordeaux, LaBRI, UMR 5800, F-33400 Talence, France

Aurélie Bugeau

University of Bordeaux, LaBRI, UMR 5800, F-33400 Talence, France

Jenny Benois-Pineau

University of Bordeaux, LaBRI, UMR 5800, F-33400 Talence, France

CONTENTS

1.1 Introduction 5
1.2 State of the Art on Objects Recognition 6
1.2.1 Features extraction 6
1.2.1.1 Global image descriptors 7
1.2.1.2 Local image descriptors 7
1.2.1.3 Semi-local image descriptors 8
1.2.1.4 Bag-of-Visual-Words approaches 9
1.2.1.5 Improvements of Bag-of-Visual-Words approaches 10
1.2.1.6 Conclusion 12
1.2.2 Classification and recognition of objects in images 12
1.2.2.1 Vector distances 13
1.2.2.2 Feature distribution comparison 13
1.2.2.3 Image classification 14
1.2.3 Object recognition in videos 14
1.3 Visual Saliency in Visual Object Extraction 14
1.3.1 State of the art in visual saliency for object extraction 15
1.3.2 Subjective vs. objective saliency maps 16
1.3.3 Objective saliency map 19
1.3.3.1 Spatial saliency map 20
1.3.3.2 Temporal saliency map 21
1.3.3.3 Geometric saliency map 22
1.3.3.4 Saliency map fusion 23
1.4 Object recognition with saliency weighting 24
1.5 Evaluation 24
1.5.1 IMMED and ADL video databases 25
1.5.2 Eye-tracker experiment 26
1.5.3 Saliency maps evaluation 26
1.5.4 Object recognition evaluation 27
1.6 Results 27
1.6.1 Saliency model assessment 28

1.6.2 BoVW vs saliency-based BoVW 28
1.6.2.1 IMMED Corpus 28
1.6.2.2 ADL Data-set 29
1.7 Conclusion 36

1.1 Introduction

Object recognition or classification have sparked the interest of researchers for nearly three decades. Nowadays, this topic is one of the most active in the computer vision research community. Object recognition/classification are performed on several digital media such as pictures or videos. The recognition task is more or less obvious according to the visual scene complexity, and the object to find. It is indeed easier to find an object in a controlled environment than in a natural scene. Furthermore, in a real-life visual scene, objects can be numerous and located in the foreground, and the background as well. The object recognition task is often dependent on the global semantic interpretation task. One do not seek to recognize all objects in a visual scene, but only those which are if interest for him. The examples of such a *selective* interest are numerous. When seeking for identifying a person crossing the road, the observer will not focus on the surrounding buildings for instance.

The book chapter addresses recognition/classification of objects in complex visual scenes recoded by using a wearable video camera. Especially we are willing to recognize manipulated objects of the *Instrumental Activities of Daily Living (IADL)*. For such videos, the wearable camera is either set on the subject's shoulder or tied on the chest. Both camera positions give an *egocentric* point-of-view of the visual scene. This point-of-view has the advantage to be the best to catch the action happening. However, nobody is behind the camera to center the object of interest. That is why the object of interest may be located in an unexpected area of the video frame. This issue is not usual in edited videos where objects of interest are almost always near the frame center. IADL video scene are complex as well. Indeed several manipulated objects could be present on the frame, but only one or two of them could be *active* that is of interest for the observer. Hence an additional information must be integrated in the recognition framework to catch the attention of the observer.

In this work, we propose to use the visual saliency for detecting *active* regions of the frame. The visual saliency represents the human visual attention within a visual scene. Therefore the saliency is well suited to distinguish active from inactive objects. Visual saliency modeling captivates researchers since the early 80's with the *Feature Integration Theory* [51] from A. Treisman, and G. Gelade. This research topic is still very active. In 2012, A. Borji, and L. Itti [8] took the inventory of 48 significant saliency models. Despite the fact that the visual saliency modeling is an old research topic, object recognition frameworks using such models is a new trend [18, 53]. Most of the visual saliency models are only considering spatial information such as contrast. These models are called *spatial* and where designed at first for still pictures. There are also models called *spatio-temporal* based on the motion present in videos. Especially the Human Visual System (HVS) is highly sensitive to the

relative motion. This is why applying a *spatio-temporal* saliency model in the object recognition framework is relevant to consider the temporal dimension of videos. Indeed most of the object recognition frameworks for video only process video frames separately, without taking advantage of previous and next frames.

In this chapter, we also propose to improve the saliency model by adding a third saliency cue called *geometric*. Recent works [50] have shown that subjects tend to fixate the screen center when watching natural scene. In [17] the authors came to the same conclusion for natural edited videos. This is why the authors of [13] proposed a thrid cue modeling a 2D Gaussian centered in the middle of the frame. In our third cue, we considered this center hypothesis applied to egocentric videos. After analyzing gaze fixations on these videos we figured out that viewers anticipate the camera motion. Hence, we propose with the *geometric* cue to consider the anticipation phenomenon by moving the 2D Gaussian center according to the camera center motion.

Before going into further details on the use of saliency for object retrieval, we review in section 1.2 the existing methods for object recognition in images or video frames considered as stills. Then the section 1.3 presents the state-of-the-art methods using visual saliency for object recognition as well as saliency models that are used in this work, and the proposed saliency *geometric* saliency cue. Section 1.4 describes our object extraction approach based on the BoVW weighted by saliency maps. Section 1.5 details the evaluation protocol, and the test video databases. Section 1.6 shows the evaluation results. Finally, section 1.7 concludes this chapter.

1.2 State of the Art on Objects Recognition

Object recognition or classification are very active research topics. Over thousands of papers have been published on these subjects during the last ten years. Doing an exhaustive state of the art is therefore unrealistic. Hence we focus on the works that have received the most attention and have given the most promising results. One common strategy for all these methods can be highlighted. First, the image or areas of interest is described with the most possible pertinent information. The descriptors can either be local, global or semi-local. Next, a compact representation of the set of all the descriptors is defined. Finally, distances or similarities between these representations are computed so that the current image can be classified or compared to a database in order to obtain the recognition result. In this section, all these steps are detailed.

1.2.1 Features extraction

In order to analyze the content of images or videos, the first step consists in extracting some features which characterize the data. This step is useful for all the applications such as Content-Based Image Retrieval (CBIR), image classification, object recognition or scene understanding. The features can either be global, local or semi-local. All of them can be applied for object recognition in the areas of interest detected in video frames. We here review some of the existing works on the topic.

1.2.1.1 Global image descriptors

Global image features are generally based on color cues. Indeed, color is an important part of the human visual perception. In images, the colors are encoded in color spaces. A color space is a mathematical model that enables the representation of colors, usually as a tuple of color components. There exist several models of this type, some motivated by the application background, some by the perceptual background of the human vision system. Among them we can cite the RGB (Red Green Blue) space, the HSV (Hue Saturation Value) or the luminance-chrominance spaces (YUV for instance).

Probably the most famous global color descriptor is the color histogram. Color histograms aim at representing the distribution of colors within the image or a region of the image. Each bin of a histogram represents the frequency of a color value within this area. It usually relies on a quantization of the color values, which may differ from one color channel to another. Histograms are invariant under geometrical transformations of the region.

Color moments are another way of representing the color distribution of an image or a region of an image. The first order moment is the mean which provides the average value of the pixels of the image. The standard deviation is the second order moment representing how far color values of the distribution are spread out from each other. The third order moment, named skewness, can capture the asymmetry degree of the distribution. It will be null if the distribution is centered on the mean. Using color moments, a color distribution can be represented in a very compact way [26, 34].

Other color descriptors that can be mentioned are the Dominant Color Descriptor (DCD) introduced in the MPEG-7 standard [37] or the Color Layout Descriptor (CLD).

1.2.1.2 Local image descriptors

The features that have received the most attention in the recent years are the local features. The main idea is to focus on the areas containing the most discriminative information. In particular the descriptors are generally computed around the interest points of the image and are therefore often associated to an interest point detector.

SIFT

Scale Invariant Feature Transform (SIFT) [35] has been designed to match different images or objects of a scene. The features are invariant to image scaling and rotation, and partially invariant to change in illumination and 3D camera viewpoint. They are well localized in both the spatial and frequency domains, reducing the probability of disruption by occlusion, clutter, or noise. In addition, the features are highly distinctive, which allows a single feature to be correctly matched with high probability against a large database of features, providing a basis for object and scene recognition. There are two main steps for extracting SIFT features: the key-point localization through scale-space extreme detection and the generation of key-point descriptors. First, a scale pyramid is built by convolving the image with variable-scale Gaussians and DoG images are computed from the difference of adjacent blurred images. Interest points for SIFT features finally correspond to local extrema of these DoG images. To determine the key-point orientation, necessary for rotation invariance, a gradient orientation histogram is computed in the neighborhood of the key-point. The contribution of each neighboring pixels is weighted by the gradient magnitude. Peaks in the histogram indicate the dominant orientations. The feature descriptor finally corresponds to a set of orientation histograms, relative to the key-point orientation, on a 4x4 pixel neighborhoods. As histograms contain 8 bins, a SIFT features is a vector of 128 dimensions. This vector is normalized to ensure invariance to illumination changes.

SURF

SIFT have proven to be a powerful feature in many computer vision applications. Nevertheless, all the necessary convolutions make it computationally expensive. Speeded Up Robust Features (SURF) [5] have then been proposed as an alternative feature. This feature describes a distribution of Haar-wavelet responses within interest point neighborhood. It relies on integral images. The latter is the sum of all pixel values contained in the rectangle between the origin and the current position. SURF key-points are also extracted by scale-space analysis through the use of Hessian-matrices. Here again, the dominant orientation is extracted. It is estimated by computing the sum of Haar-wavelet responses within a sliding orientation window. In an oriented square window centered at the key-point, which is split up into 4x4 sub-regions, each sub-region finally yields a feature vector based on the Haar-wavelet responses, of dimension 64.

1.2.1.3 Semi-local image descriptors

Most shape descriptors fall into this category. Shape description relies on the extraction of accurate contours of shapes within the image or region of interest. Image segmentation is usually fulfilled as a preprocessing stage. In order for the descriptor to be robust with regard to affine transformations of

an object, quasi perfect segmentation of shapes of interest is supposed. Here, we just mention some shape descriptors but more can be found in literature. In particular, let us mention the Curvature Scale Space (CSS) descriptor [39] and the Angular Radial Transform (ART), descriptors in the MPEG-7 standard.

1.2.1.4 Bag-of-Visual-Words approaches

The descriptors presented above, and in particular SIFT and SURF, have been widely used for retrieving objects in images. Local feature extraction leads to a set of unordered feature vectors. The main difficulty of the recognition, retrieval or classification steps consists in finding a compact representation of all these features and its associated (dis-)similarity measure. An efficient approach that has been widely used is the so-called Bag-Of-Visual-Words framework [47], that we now describe. The Bag-of-Visual-Words (BoVW) approaches have four main stages: building a visual dictionary by clustering visual features extracted from a training set of images/objects, quantifying the features, choosing an image representation using the dictionary and comparing images according to this representation. We now review these steps.

Visual dictionary

In analogy with text retrieval, the features extracted in an image correspond to the words in a document. A visual dictionary must then be built. This is generally done by randomly selecting a sufficiently large set of features over a huge amount of images. This dictionary, $V = v_i, i = \{1, \ldots, K\}$, is then built by clustering these features into a certain number of K classes or "visual words".

Feature quantization

The second step consists in quantizing the features extracted in an image according to the visual dictionary. Each feature from N extracted features for an image is *quantized*. This quantization is generally achieved by assigning each feature to its closest word in the dictionary V.

Pooling

Each image in the data-set can now be represented by a unique vector of K dimensions. Each dimension represents the number of times a feature appears in the image. Therefore, this vector can be seen as a histogram representing the distribution of visual words in an image. This histogram is often normalized which enables comparing images containing a different number of features. These histograms were named *Bag-of-Visual-Words* [47] *(BoVW)*.

Image comparison

All images being now represented by a histogram, the last step simply consists in comparing the histograms. Obviously, when the size of the database increases this step can become very computationally expensive. The computational time also depends on the size of the dictionary which therefore needs to be chosen carefully. Several strategies have been proposed in the literature to improve the cost of this last step. In [47], this framework was applied with SIFT features. The vector quantization was carried out by k-means clustering, the number of clusters being chosen manually. In order to increase the discriminative power of the words produced by the clustering, a stop list was used. The idea of the stop list is to remove from the vocabulary the words which are very frequent, thus not discriminative enough, and those which are very rare, that can then be seen as noise. In [47], the authors removed from the list the 5% more frequent words and 10% less frequent.

Limitations and improvements

The method in [47] is at the origin of most recent works in the domain of image recognition and retrieval. Many improvements have been proposed since then.

1.2.1.5 Improvements of Bag-of-Visual-Words approaches

Feature quantization

First, concerning the vector quantization, it is well known that k-means algorithm has no guarantee to converge to the global optimum and depends on the initialization of the centers of the clusters. An improved version of this algorithm, known as k-means++ has been proposed in [2]. In order to deal with an incremental amount of images in a data-set, a hierarchical quantization can be built. For instance, a hierarchical k-means clustering, called vocabulary tree was proposed in [40]. The vocabulary tree gives both higher retrieval quality and efficiency compared to the initial BoVW framework of [47]. Until now, we have only been talking about Bag-of-Visual Words approaches in which only one type of feature is used. Note that if several different types of features are extracted from the images, the BoVW framework can also be directly applied. The set of all the feature vectors from one image is generally referred to as "Bag-of-Features".

Soft and sparse coding

In previous methods, the feature quantization, and thus the image representation, is obtained by assigning the feature vector to the closest word in the dictionary. This is called "hard coding". The coding step can be modeled by a function which assigns a weight $\alpha_{i,j}$ to the closest center v_j of the feature

vector x_i :

$$\alpha_{i,j} = \begin{cases} 1 & \text{if } \mathrm{j} = \arg\min_{\mathrm{K}} \|\mathrm{x_i} - \mathrm{v_K}\| \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

Hence for each feature vector in the image, a code-vector wan be computed by encoding the feature with the dictionary. The drawbacks of hard quantization are twofold: (i) word uncertainty: when a feature is close to several codewords, i.e. words of the dictionary, only the closest is considered; (ii) word plausibility: a codeword is assigned to the closest codeword no matter how far it can be. An illustration of these two drawbacks of hard coding are given in Figure 1.1 below. In green an example of an "uncertain" word is shown, in light blue an example of an implausible word is given. A "consistent" word example is given in yellow.

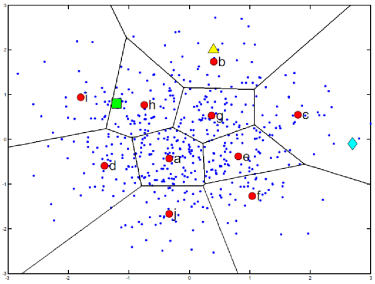


FIGURE 1.1
Soft quantization and sparse coding.

Instead of assigning a feature to a unique codeword, a soft assignment can be used [52]. The weight α is not anymore a constant weight for all the features vectors but will contain more information on the distribution of the feature vectors over the visual dictionary. For instance, the weight can be equal to the distance from the feature to the codeword. The resulting code is therefore not sparse contrary to what happens with hard coding. Some works also proposed to use kernel density estimators to estimate the weights [7]. Let us mention three models studied in [7], and relying on a Gaussian-shaped kernel: the kernel code-book (KCB), the codeword uncertainty (UNC) and the codeword plausibility (PLA). They all weight each word by the average kernel density estimation for each feature.

Sparse coding uses a linear combination of a small number of codewords to approximate a feature. The strength of sparse coding approaches is that one can learn the optimal dictionary while computing the weights α [36].

Pooling in BoVW approaches

The third step of BoVW approaches is pooling which consists in forming the final image representation. A good representation must be robust to different image transformations and to noise, and must be as compact as possible. A pooling operator aggregates the projections of all the input feature vectors onto the visual dictionary to get a single scalar value for each codeword. The standard BoVW [47], considers the traditional text retrieval sum pooling operator:

$$\forall j = 1 \dots K \ , \ z_i = \sum_{i=1}^{N} \alpha_{i,j}.$$

Max pooling,

$$\forall j = 1 \dots K \ , \ z_i = \max_{i=1,\dots,N} \alpha_{i,j} \ ,$$

associated to sparse coding have also allowed to get superior performance than sum pooling [55]. Performance of max pooling and sum pooling has also been studied in [12]. Several extensions to these two traditional pooling operators have recently been proposed, some focusing on applying the pooling step on more local areas. The most powerful is probably the Spatial Pyramid Matching method (SPM) [31]. A fixed predetermined spatial image pyramid is first computed. The BoVW are then built on nested partitions of image plane from coarse-to-fine resolutions. In other words, pooling is performed over the cells of a spatial pyramid rather than over the whole image. In [1], an approach called "Visual Phrases" is introduced to group visual words according to their proximity in the image plane as a sequence of features. The visual phrases are represented by a histogram containing the distribution of the visual words in the phrase. Spatial information has also been taken into account in [27]. Indeed, a spatial embedding of features with local Delaunay graphs is proposed. The advantage of Delaunay triangulation is that it is invariant to affine transformation of image plane preserving the angles. Another BoVW improvement belonging to the aggregated coding class is the Fisher Kernel approach proposed in [41]. It is based on the use of the Fisher kernel with Gaussian Mixture Models (GMM) estimated over the whole set of images.

1.2.1.6 Conclusion

In this section, we gave an overview of different type of features to represent the images and videos. The extraction of such descriptors is the preliminary step for any visual indexing and retrieval systems. The accuracy of all the methods presented in the following highly depends on the robustness of the chosen features to scaling, rotation and lightning changes.

1.2.2 Classification and recognition of objects in images

Now that a good representation of each image or video has been extracted, the problem of classification or retrieval can be addressed. In case of image retrieval or indexing, the goal is to find, within a database, the image(s) that best matches a query image given by the user. In the context of classification, the purpose is to assign the image to the category to which it corresponds. The categories are defined beforehand by the user and a learning phase is necessary to learn the most important properties of each category. When relying to BoVW approaches, at the end of the pooling step, every object or image is represented by one histogram over the visual dictionary. In this section, we will see how these histograms can be used for image or object retrieval on one hand and for classification on the other hand.

1.2.2.1 Vector distances

Many distances and strategies have been proposed to retrieve an image from its compact representation. Obviously, as BoVW approaches represent the distribution of visual words in an image by a histogram, the easiest way to perform retrieval is to compute (dis-)similarities between histograms. There exist two main categories of distances between histograms: the bin-to-bin distances and the cross-bin ones. Bin to bins measures require the histograms to have the same number of bins. Among the existing ones, let us mention the L2 and L1 metrics, the Kullback-Leibler divergence, the Chi-Square metric or the histogram intersection [49]. Among the cross-bins metrics, the most famous ones are the Mahalanobis distance and the Earth Mover's distance (EMD).

1.2.2.2 Feature distribution comparison

In previous sections, we consider a unique histogram per image. However, more recent approaches have shown that it can be more powerful to represent an image by several histograms, taking into account only a small part of the data or incorporating some local descriptions. When the data is represented by one or several histograms, kernel functions can be used to perform the matching. A kernel function is a function which allows evaluating the correlation between two data descriptions. In recent years, two principal types of kernels have been used in visual recognition system: the pyramid match kernel and the context dependent kernel. Pyramid match kernel was introduced in [20], for object recognition and document analysis. The principle is to map the features of some interest points using a multi-resolution histogram representation and to compute the similarity using a weighted histogram intersection. In [45], Sahbi et al. have introduced the context dependent kernel (CDK). This kernel takes into account both the feature similarity "alignment quality" and the spatial alignment in a neighborhood criterion. The CDK is defined as the fixed-point of an energy function which balances a "fidelity" term, i.e. the

alignment quality in terms of features similarity, a context criterion, i.e. the neighborhood spatial coherence of the alignment and an entropy term.

1.2.2.3 Image classification

Image classification requires a pre-processing step of learning in order to find a decision rule (classifier) assigning Bag-of-Features representations of images to different classes. In general, a set of images belonging to the class (positive training examples) and a set of images not belonging to the class (negative training examples) are provided to the learning tool. Nowadays, SVM is the most frequently used machine learning tool in a supervised context. Hence we find it necessary to briefly review its principles for object recognition purposes. SVM is a supervised statistical learning method that belongs to the class of kernel methods. The goal of SVM is to learn good separating hyper planes in a high dimensional feature space. The role of the kernel function k is to map the training data into a higher dimensional space where the data is linearly separable. In other words, the feature vectors, x_i, are first mapped into feature vectors $\phi(x_i)$ in an induced space. Next, a linear decision function $f(x_i) = wx_i + b$ is defined. The hyperplane then corresponds to $wx_i + b = 0$ and separates the positive $y_i = +1$, from the negative training examples, $y_i = -1$:

$$\begin{cases} wx_i + b \geq 0 & \text{for } \mathrm{y_i} = +1 \\ wx_i + b < 0 & \text{for } \mathrm{y_i} = -1. \end{cases} \tag{1.2}$$

The SVM optimization problem can now be formulated as

$$min_{w,b} \frac{1}{2}\|w\|^2 \text{ subject to } \forall \mathrm{i},\ \mathrm{y_i f(x_i)} \geq 1. \tag{1.3}$$

The decision function relies on the so-called "support vectors" which define the maximal margin between positive and negative subspaces in the target space. Detailed description can be found in the fundamental work []. In its original formulation SVM is a binary classifier, but since the original framework it has been adapted to the multi-class problem.

1.2.3 Object recognition in videos

For recognition of objects in videos, a lot of work has been done using so-called spatio-temporal features [46] computed *around* spatio-temporal points [30]. Nevertheless, the key-framing and intra-frame object recognition still remains one of the most popular approach [16]. Temporal dimension can be integrated in this case by fusion operators using multiple detections along the video [4] or by extraction of visually salient regions which are supposed to contain object of interest. In the following section we will introduce the notions of visual saliency for object extraction in videos.

1.3 Visual Saliency in Visual Object Extraction

1.3.1 State of the art in visual saliency for object extraction

Since recently, the focus of attention in video content understanding, presentation, and assessment has moved towards incorporating of visual saliency information to drive local analysis process. The fundamental model by L. Itti and C. Koch [24] is that one the most frequently used for driving analysis process by visual saliency. If we simplify the concept of saliency to its very basic definition, we can reasonably say that visual saliency is what attracts human gaze. For visual object recognition in video, the first step which consists in extracting the potential area of object can be driven by extraction of saliency areas in video frames. Then features can be selected in these areas for object description. Numerous psycho-visual studies which have been conducted since the last quarter of 20th century uncovered some factors influencing it. Considering only signal features, the sensitivity to color contrasts, contours, orientation and motion observed in image plane has been stated by numerous authors [23, 32]. Nevertheless, only these features are not sufficient to delimit the area in the image plane which is the strongest gaze attractor. In [50], the author states, for still images, that observers show a marked tendency to fixate the center of the screen when viewing scenes on computer monitors. The authors of [17], come to the same conclusion for dynamic general video content such as movies and Hollywood trailers. This is why the authors of [13] propose the third cue which is the geometrical saliency modeled by a 2D Gaussian located at the image center. While signal based cues remain valuable saliency indicators, we claim that geometrical saliency depends on global motion and camera settings in the dynamic scene. Recently a new and challenging video content came out from various applications: the so-called *egocentric vision* content [33, 48]. It is recorded by camera worn by persons. This is a complex content, characterized by a strong camera motion, richness of the visual scene, especially if recorded in a home environment as it is done in [27] for studies of neurodegenerative diseases. Object recognition in such videos is difficult due to occlusions by hands, and the complexity of the environment. This is probably the most challenging content from the variety of user generated content from mobile devices. Hence the methods for visual object recognition in such content would make profit to the advances in a wide range of visual object recognition tasks in video. Some attempts to identify visual saliency for object recognition mainly on the basis of the frequency of repetition of visual objects and regions in the wearable video content have recently been made in [44]. We are specifically interested in the development and the application of visual saliency extraction methods for object recognition inhere. We consider visual saliency as a combination of all cues in the pixel domain for the case of "egocentric" video content.

Recent research for object recognition extend the Bag-of-Features video

representations [46] by making use of space-variant saliency mask [53]. In [46], space-time descriptors, capturing spatial appearance and motion properties, are sampled densely over the entire scene. The authors of [53] then propose to prune this set of densely extracted descriptors with the cumulative distribution function:

$$F(x;k;\lambda) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k}, \tag{1.4}$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. The raw saliency value $x \in [0, 1]$ is derived from a saliency mask obtained with six different saliency models.

In the following section, we propose an automatic method for spatio-temporal saliency extraction on wearable camera videos with a specific accent on geometrical saliency dependent on strong wearable camera motion. We evaluate the proposed method with regard to subjective saliency maps obtained from gaze tracking. The obtained saliency maps will serve for weighting the visual features in the whole BoVW video object recognition scheme we use in this work. The advantage of our approach against [53] is that we propose a trully automatic way of building saliency maps. Specifically, for the "egocentric vision" the research on visual saliency is in its embryonic stage.

Let us now introduce the notion of visual saliences, the one *subjective* we can obtain from human observers of the video content, and other *objectives*, that are automatically predicted from the video signal features.

1.3.2 Subjective vs. objective saliency maps

Any *objective* human visual perception model expressed as a induced visual attention map in the image/video plan has to be validated and evaluated with regard to a *ground truth*. The *ground truth* is the *subjective* saliency. The subjective saliency is built from eye fixation measurements with the help of eye-tracking. The eye positions are recoded with a device called eye-tracker. The eye-tracker only collects eye positions at a regular rate, up to 1250 Hz for some models. Eye positions are first measured in the eye-tracker coordinates system. Then the measures are transposed to the experiment screen coordinates system, and recorded. The origin of the screen coordinates system is usually the screen center. Finally, the measures have to be transposed to the frame coordinates system. In this chapter, we consider that the eye measure coordinates (x_0, y_0) are already transposed in the frame coordinates system. However eye fixations cannot be directly used to represent the visual attention. First the eye fixations are only spots on the frame and do not represent the field of view. Secondly, to get accurate results, the saliency map is not built with the eye tracking data from one subject, but from many subjects. So the subjective saliency map should provide an information about the density of eye positions.

The method proposed by D. S. Wooding [54] fulfills these two constraints. Moreover, his method was tested over 5000 participants on digitalized images

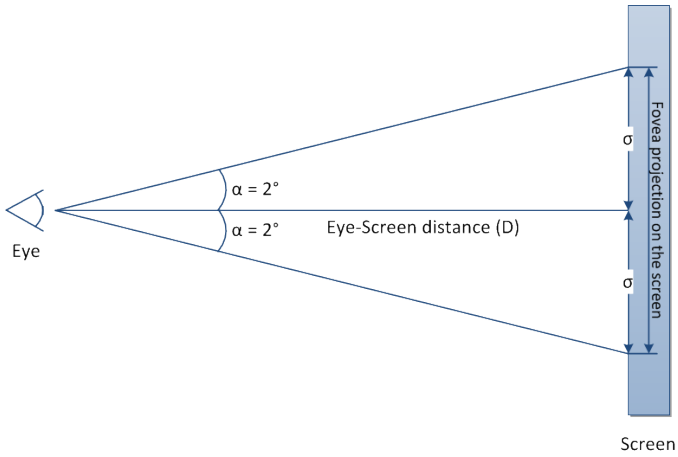


FIGURE 1.2
Fovea projection on the screen

of paintings from the National Gallery. In the case of video sequences, the method is applied on each frame I of a sequence M. The process result is a subjective saliency map $S_{subj}(I)$ for each frame I. With this method, the saliency map is computed in three steps. In the first step, for each eye fixation measure m of frame I, a two dimensional Gaussian is applied at the center of the eye measure $(x_0, y_0)_m$. The two dimensional Gaussian depicts the fovea projection on the screen. The fovea is the central retina part where the vision is the most accurate. The image falls on the fovea when an observer fixates. This region contains only cone photo-receptors and has the highest cone density of the retina. The human eye contains two kinds of receptors, the rods and the cones. Rods are more sensitive at low light levels. However, rods do not allow color discrimination and provide poor information about details. On the contrary, cones are efficient at high intensity lightning. Cones are responsible for color vision and for the fine details detection. In the *Sensibility to Light* [21] book chapter from D.C. Hood and M.A. Finkelstein (1986), the authors stated that the fovea covers an area from 1.5° to 2° in diameter at the retina center. It is also specified that the cone population falls sharply outside the fovea region, to reach a minimum at around 10° from the fovea center. M. Pomplun, H. Ritter and B. Velichkovsky in [43] were the first to apply a two dimensional Gaussian to depict the fovea projection on the screen. D.S. Wooding proposed to set the Gaussian spread σ to an angle of 2° (Figure 1.2). Equation (1.5) is used to estimate the σ_{mm} in mm according to the fovea view angle α set to 2°. The distance eye-screen D in mm is also required. According to *ITU-R Rec. BT.500-11* [25], D should be equal to three times the screen height $3H$.

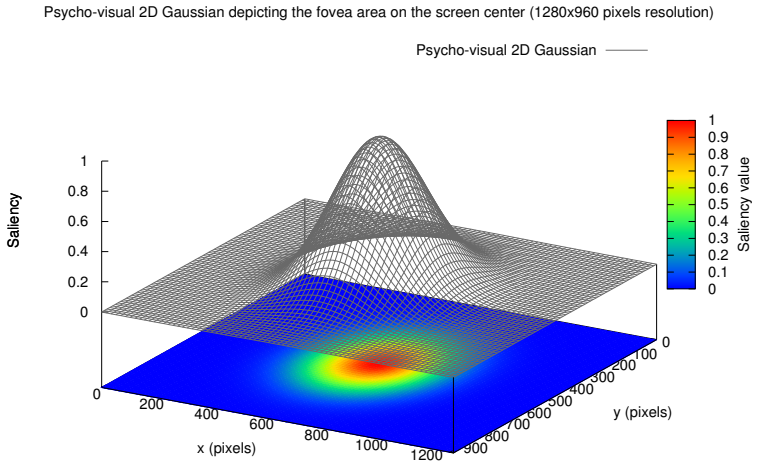


FIGURE 1.3
Psycho-visual 2D Gaussian (2° spread) depicting the fovea area related to one eye-tracker measure.

$$\sigma - mm = D \times tan\,(\alpha) \tag{1.5}$$

Nevertheless, measures in mm are not convenient for processing video frames. So the σ_{mm} value in mm is converted in pixels (σ) with equation (1.6), where R is the screen resolution in pixels per mm .

$$\sigma = R \times \sigma_{mm} \tag{1.6}$$

For the eye measure m of the frame I, a partial saliency map $S_{subj}(I, m)$ is computed (1.7).

$$S_{subj}(I, m) = Ae^{-\left(\frac{(x-x_{0_m})^2}{2\sigma_x^2}+\frac{(y-y_{0_m})^2}{2\sigma_y^2}\right)}$$
$$\text{with } \sigma_x = \sigma_y = \sigma \text{ and } A = 1 \tag{1.7}$$

Then, at the second step, all the partial saliency maps $S_{subj}(I, m)$ of frame $S_i I$ are added into $S_{subj}\prime(I)$ (1.8).

$$S_{subj}\prime(I) = \sum_{m=0}^{N_I} S_{subj}(I, m) \tag{1.8}$$

where N_I is the number of eye measures recorded on all the subjects for the frame I. Finally, at the third step, the saliency map $S_{subj}\prime(I)$ is normalized by

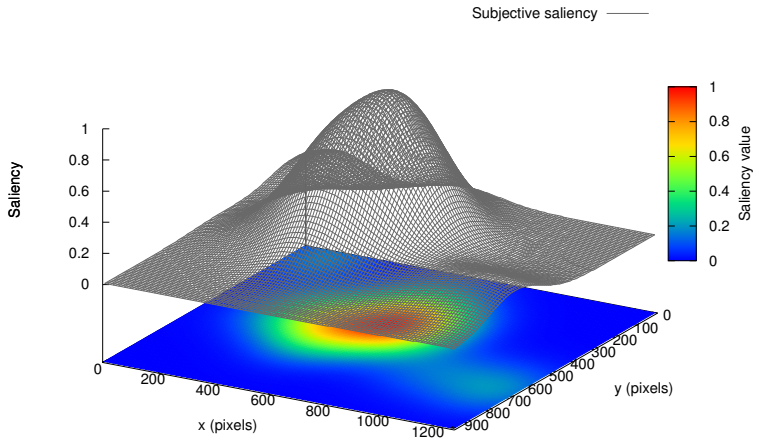
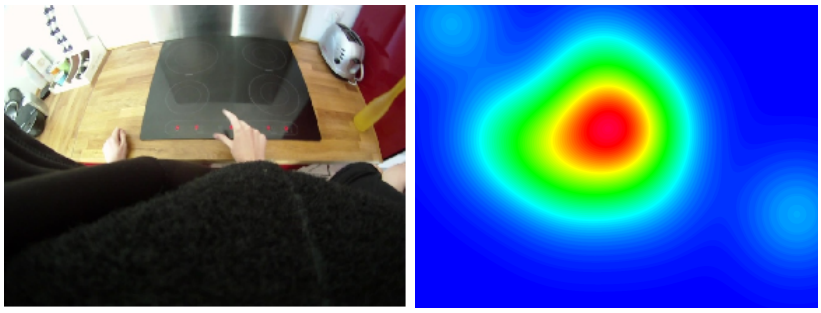


FIGURE 1.4
Normalized 2D Gaussian sum on an egocentric video frame.

the highest value $argmax$ of $S_{subj}\prime(I)$ (Figure 1.4). The normalized subjective saliency map is stored in $S_{subj}(I)$ (1.9).

$$S_{subj}(I) = \frac{S_{subj}\prime(I)}{argmax(S_{subj}\prime(I))} \tag{1.9}$$

Figure 1.5 shows an example of a subjective saliency map computed with the D. S. Wooding's method. Why could not we use the subjective saliency maps for object extraction? The *subjective* saliency map processing requires eye-tracker measures from subjective experiments. Subjective experiments are time consuming and expensive to carry out. To get a *subjective* saliency map would require viewing the video by several subjects. Thus, *subjective* saliency are not suited for real-life video analysis applications. To fulfill this constraint we are interested in an automatic *objective* saliency maps we proposed. The automatic spatio-temporal saliency map computation process is described in the next sections. However the *subjective* saliency map remains helpful for the objective saliency map accuracy evaluation. That is why the *subjective* saliency map is considered below as the reference saliency map. Hence, the *subjective* saliency maps will be used as the ground truth to asses the objective maps automatically built.



(a) (b)

FIGURE 1.5
Subjective saliency map example computed with D. S. Woodings method from eye-tracker data. (a) Original frame. (b) Subjective saliency map.

1.3.3 Objective saliency map

To delimit the area of video analysis in video frames to the regions which are potentially interesting to human observers we need to model visual saliency on the basis of video signal features. Here we follow the results of community research [22, 24, 23, 32, 13, 44] by proposing fusion of spatial, temporal, and geometric cues. We extend the state-of-the art approaches by a specific modeling of geometrical saliency and propose multiplicative fusion of all three cues.

(a) (b) (c)

FIGURE 1.6
Objective saliency map example. (a) Original frame. (b) Spatial saliency map. (c) Temporal saliency map.

1.3.3.1 Spatial saliency map

The spatial saliency map S_{sp} is mainly based on color contrasts [3]. We used the method from O. Brouard, V. Ricordel and D. Barba [13]. The spatial saliency map extraction is based on seven color contrast descriptors. These descriptors are computed in the HSI color space [19]. On the contrary to RGB color system, the HSI color space is well suited to describe color interpretation by humans. The spatial saliency is defined according to the following seven local color contrasts V in the HSI domain:

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 $S'_{sp}(I,i)$ for a pixel i from a frame I is computed by mean fusion operator from seven color contrast descriptors (1.10):

$$S'_{sp}(I,i) = \frac{1}{7}\sum_{\varsigma=1}^{7} V_{\varsigma}(I,i) \tag{1.10}$$

Finally, $S'_{sp}(I,i)$ is normalized between 0 and 1 to $S_{sp}(I,i)$ according to its maximum value.

1.3.3.2 Temporal saliency map

The objective spatio-temporal saliency map model requires a temporal saliency dimension. This section will describe how to build temporal saliency maps. The temporal saliency map S_t models the attraction of attention to motion singularities in a scene. The visual attention is not grabbed by the motion itself. The gaze is attracted by the motion difference between the *absolute* motion scene and the global motion scene. The motion difference is called the residual motion. O. Le Meur et al. [32], O. Brouard et al. [13], and S. Marat [38] propose a temporal saliency map model that takes advantage

of the residual motion. In this work, we have implemented the model from O. Brouard et al. [13].

The temporal saliency map is computed in three steps. The first one is the optical flow estimation. Then the global motion is estimated in order to get the residual motion. Finally a psycho-visual filter is applied on the residual motion.

To compute the optical flow, we have applied the Lucas Kanade method from OpenCV library [9]. The optical flow was sparsely computed on 4x4 blocks, as good results were reported in [10] when using 4×4 macro-block motion vectors from the H.264 AVC compressed stream. The next step in temporal saliency computation is the global motion estimation.

The goal here is to estimate a global motion model to differentiate then local motion from camera motion. In this work, we follow the preliminary study from [10] and use a complete first order affine model (1.11):

$$\begin{aligned} dx_i &= a_1 + a_2 x + a_3 y \\ dy_i &= a_4 + a_5 x + a_6 y \end{aligned} \tag{1.11}$$

Here $\theta = (a_1, a_2, \ldots, a_6)^T$ is the parameter vector of the global model (1.11) and $(dx_i, dy_i)^T$ is the motion vector of a block. To estimate this model, we used robust least square estimator presented in [28]. We denote this motion vector $\vec{V}_\theta(I, i)$. Our goal is now to extract the local motion in video frames i.e. residual motion with regard to model (1.11). We denote the macro-block optical flow motion vector $\vec{V}_c(I, i)$. The residual motion $\vec{V}_r(I, i)$ is computed as a difference between block motion vectors and estimated global motion vectors.

Finally, the temporal saliency map $S_t(I, i)$ is computed by filtering the amount of residual motion in the frame. The authors of [13] reported, as established by S. Daly, that the human eye cannot follow objects with a velocity higher than $80°/s$ [15]. In this case, the saliency is null. S. Daly has also demonstrated that the saliency reaches its maximum with motion values between $6°/s$ and $30°/s$. According to this psycho-visual constraints, the filter proposed in [13] is given by (1.12).

$$S_t(s_i) = \begin{cases} \frac{1}{6}\vec{V}_r(I, i), & \text{if } 0 \leq \vec{V}_r(I, i) < \vec{v}_1 \\ 1, & \text{if } \vec{v}_1 \leq \vec{V}_r(I, i) < \vec{v}_2 \\ -\frac{1}{50}\vec{V}_r(I, i) + \frac{8}{5}, & \text{if } \vec{v}_2 \leq \vec{V}_r(I, i) < \vec{v}_{max} \\ 0, & \text{if } \vec{v}_{max} \leq \vec{V}_r(I, i) \end{cases} \tag{1.12}$$

with $\vec{v}_1 = 6°/s$, $\vec{v}_2 = 30°/s$ and $\vec{v}_{max} = 80°/s$. We follow this filtering scheme in temporal saliency map computation.

1.3.3.3 Geometric saliency map

Many studies have showed that the observers are attracted by the screen center. In [13], the geometrical saliency map is proposed as a 2D Gaussian

located at the screen center with a spread $\sigma_x = \sigma_y = 5°$. In our work [11] we proposed to adapt the geometric saliency to the camera position estimated by a psycho-visual experiment with subjects watching recorded videos. We stated that in a shoulder-fixed wearable camera video the gaze is always located in the first upper third of video frames, see the scattered plot of subjective saliency peaks in Fig. 1.7. Therefore, we have set the 2D Gaussian center at $x_0 = \frac{width}{2}$ and $y_0 = \frac{height}{3}$. The geometrical saliency S_g map equation is given by (1.13).

$$S_g(I) = e^{-\left(\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2}\right)} \tag{1.13}$$

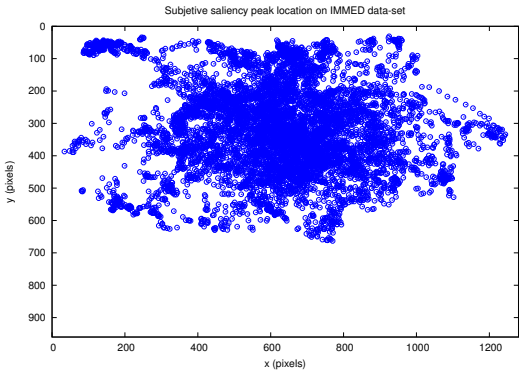


FIGURE 1.7
Scattered plot of subjective saliency peaks for all the frames from the IMMED database

However, this attraction may change with the camera motion. This is explained by the anticipation phenomenon [29]. 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, equation (1.11), where $x = \frac{width}{2}$ and $y = \frac{height}{3}$.

1.3.3.4 Saliency map fusion

We now describe the method that merges these three saliency maps in the target *objective* saliency map. The fusion result is a spatio-temporal-geometric saliency map. In [10], several fusion methods for the spatio-temporal saliency without geometric component were proposed. We have tested these fusion

methods on wearable video database. The results show that the multiplicative fusion performs the best. So for the full spatio-temporal-geometric saliency we compute multiplicative S^{mul}_{sp-t-g} (1.14).

$$S^{mul}_{sp-t-g}(I) = S_{sp}(I) \times S_t(I) \times S_g(I) \tag{1.14}$$

1.4 Object recognition with saliency weighting

In the trending approach of Bag-of-Visual-Words, all the SURF descriptors from the learning dataset are quantized into a visual dictionary (or codebook) using *k-means* clustering. In this case, SURF points are detected by using the sparse SURF detector. Each image is then modeled by a distribution of its visual words. For this purpose, the descriptors computed on the image are matched with a L_2 norm with their closest representative in the codebook. In the traditional Bag-of-Visual-Words approach, the final image signature is the statistical distribution of the image descriptors according to the codebook. We propose instead of a hard assignment for BoVW computation (see section 1.2.1.4) to apply what we call *saliency weighting*. With saliency weighting, the contribution of each image descriptor is defined by the maximum saliency value found under the descriptor φ_i. In other words, descriptors over salient areas will get more weight in the image signature than descriptors over non-salient areas, see Figure 1.8. Therefore, the image signature z_i is computed with equations (1.15), and (1.16):

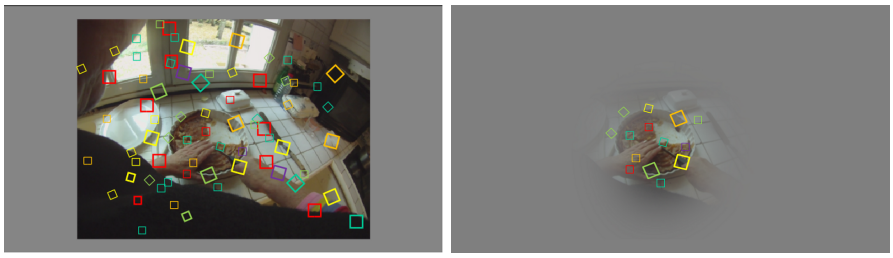
$$\forall j = 1 \dots N, z_i = \sum_{i=1}^{N} \alpha_{i,j} w_i \tag{1.15}$$

$$w_i = \underset{s \in \varphi_i}{argmax}(S(s)) \tag{1.16}$$

where $\alpha_{i,j} = 1$ if the descriptor φ_i of feature point p_i is quantized to class K. w_i is the maximum saliency map value within the area covered by the descriptor φ_i.

1.5 Evaluation

This section presents the video databases and the methods used for the evaluation of the saliency model and the object recognition. Recognition methods



(a) (b)

FIGURE 1.8
Example of SURF points not-weighted, and weighted by visual saliency. (a) Classical BoVW. (b) BoVW weighted by visual saliency.

were tested on the IMMED[1] (Indexing MultiMEdia data from wearable sensors for diagnostics and treatment of Dementia) and the ADL (Activities of Daily Life) [42] video databases. Both are egocentric video corpus depicting Instrumental Activities of Daily Living (IADL). The evaluation of saliency models requires a subjective experiment to record eye fixations as presented in section 1.3.2. The duration of these experiments is limited to 30 minutes because of the tiredness of participants [25]. For this reason we only evaluate the automatic saliency models on the IMMED database. Object recognition is evaluated on both corpora. In the following subsection we describe these two data-sets.

1.5.1 IMMED and ADL video databases

The ADL dataset, and the corpus provided by the IMMED project. Both of these datasets were chosen since they were filmed by a GoPro wearable Camera which captures videos at the rate of 30 frames per second, with a resolution of 1280x960, and a 170 viewing angle.

The ADL is a publicly available academic data-set of 18 actions of daily life accomplished by 20 different people. All the 32662 frames extracted from these videos were annotated with an action label, object bounding boxes, object identity and human-object interaction.

The IMMED corpus is composed of 53 videos of activities of daily living shot in a home environment. This corpus was recorded during the time life of IMMED Project funded by French National Agency of Research (ANR). A psychologist was present during the shooting of the videos to suggest the

[1] http://immed.labri.fr

activities. A total of 3641 frames extracted from the videos were annotated with temporal tasks, object locations and object categories.

1.5.2 Eye-tracker experiment

The subjective saliency maps expressing user attention are obtained on the basis of psycho-visual experiment consisting in measuring the gaze positions on videos from wearable video camera. The map of the visual attention has to be built on each frame of these videos. Videos from wearable camera differ from traditional video scenes: the camera films the user point of view, including his hands. Unlike traditional videos, wearable camera videos have a very high temporal activity due to the strong ego-motion of the wearer. The gaze positions are recorded with an eye-tracker. We used HS-VET 250Hz from Cambridge Research Systems Ltd. This device is able to record 250 eye positions per second. The videos we display in this experiment have a frame-rate of 29.97 frames per second. A total of 28 videos from IMMED database filming the IADL of patients and healthy volunteers are displayed to each participant of the experiment. This represents 17 minutes and 30 seconds of video. The resolution of the videos is 1280x960 pixels and the storage format is raw YUV 4:2:0. The experiment conditions and the experiment room is compliant to the recommendation ITU-R BT.500-11 [25]. Videos are displayed on a 23 inches LCD monitor with a native resolution of 1920x1080 pixels. To avoid image distortions, videos are not resized to screen resolution. A mid-gray frame is inserted around the displayed video. 25 participants were gathered for this experiment, 10 women and 15 men. For 5 participants some problems occurred in the eye-tracking recording process. So we decided to exclude these 5 records. After looking at gaze position records on video frames, we stated that gaze anticipated camera motion and user actions. This phenomenon has been already reported by M. Land et al. in [29]. They state that visual fixation does precede motor manipulation, putting eye movements in the vanguard of each motor act, rather than as adjuncts to it. The observers somehow anticipate the motor act of camera wearer, in the same way as they are involved in the upcoming action. Nevertheless, gaze positions cannot directly be applied as ground truth to compare automatic saliency model we aim at. They must be processed in order to get the subjective saliency map as depicted in section 1.3.2.

1.5.3 Saliency maps evaluation

In this section, we compare the *objective* spatio-temporal saliency maps with *subjective* saliency map obtained from gaze tracking S_{subj}.

Here, we use the Normalized Scanpath Saliency (NSS) metric that was proposed in [14, 38]. The NSS is a Z-Score that expresses the divergence of the subjective saliency maps from the objective saliency maps. The NSS computation for a frame I is depicted by:

$$NSS = \frac{\overline{S_{subj} \times S_{obj}^N} - \overline{S_{obj}}}{\sigma(S_{obj})} \tag{1.17}$$

Here, S_{obj}^N denotes the objective saliency map S_{obj} normalized to have a zero mean and a unit standard deviation, $\bar{X}$ means an average. When $\overline{S_{subj} \times S_{obj}^N}$ is higher than the average objective saliency, the NSS is positive; it means that the gaze locations are inside the saliency depicted by the objective saliency map. In other words, higher the NSS is, more objective and subjective saliency maps are similar. The NSS score for a video sequence is obtained by computing the average of NSS for all frames as in [38]. Then the overall NSS score on each video database is the average NSS of all video sequences. Saliency model evaluation results are presented in section 1.6.1.

1.5.4 Object recognition evaluation

For the evaluation process we separate learning and testing images by a random selection. On each data set, 50% of the images of each category are selected as learning images for building the visual dictionaries and for the retrieval task. Here we test the standard BoVW approach using 64 dimensional SURF descriptors. For the training we applied the manually annotated masks on the object. For the testing, we compare the performance of queries by using the manually annotated mask, without mask, and the Spatio-Temporal-Geometric saliency map.

The performance is evaluated by the Mean Average Precision (MAP) measure using the *Trec Eval* [6] tool. For each test image, all images in the learning set are ranked from the closest (in terms of L_1 distance between visual signatures) to the furthest. The average precision AP aims to evaluate how well the target images, i.e images of the same class as the query, are ranked amongst the n retrieved images:

$$AP = \frac{\sum_{k=1}^{n} P(k) \cdot rel(k)}{c_p} \tag{1.18}$$

where $rel(k)$ equals 1 when the k^{th} ranked image is a target image and 0 otherwise and c_p is the total number of target. The average precision is evaluated for each test image of an object, and the MAP is the mean of these values for all the images of an object in the test set. For the whole database we measure the performance by the average value of the MAP i.e. we do not weight the MAP per class by the number of query which would give more consideration to categories where more testing images are present.

For the Spatio-temporal-geometric saliency masking we use a 2D Gaussian function located at the screen center with a spread $\sigma_x = \sigma_y = 5°$. More details regarding the evaluation of the saliency-based masking are given in the following part of this section. For these preliminary results the dictionary size has been fixed to 500, 1000, and 5000.

1.6 Results

In this section, we first report the results of the correlation of the proposed saliency method with subjective saliency. Then we present the performance of object recognition by using the ideal mask, no mask, and the spatio-temporal-geometric saliency map for query images.

1.6.1 Saliency model assessment

In this section, we compare the correlation of three automatic saliency maps with the subjective saliency. These three saliency maps are the spatio-temporal saliency map, the spatio-temporal-geometrical without camera motion, and the proposed method the spatio-temporal-geometrical with camera motion, expressing the anticipation phenomenon. The 28 video sequences described earlier from wearable cameras are all characterized by strong camera motion which is up to 50 pixels magnitude in the centre of frames. As it can be seen from the Figure 1.9 the proposed method with moving of geometrical Gaussian almost systematically outperforms the base-line spatio-temporal saliency model and the spatio-temporal-geometrical saliency with a fixed Gaussian. For few sequences (e.g. number 2), the performance is poorer than obtained by geometric saliency with a fixed Gaussian. In these visual scenes, the distractors appear in the field of view. The resulting subjective saliency map then contains multiple maxima due to the unequal perception of scenes by the subjects. This is more "semantic saliency" phenomenon (faces, etc) which can not be handled with the proposed model. The average NSS on the whole database also shows the interest of proposed moving geometrical saliency. The mean NSS scores are respectively 1.832 for patio-temporal, 2.607 for spatio-temporal with still geometrical Gaussian, and 2.791 with moving geometrical Gaussian. Which means 52.37% improvement of correspondence with subjective visual saliency map, which was our goal.

1.6.2 BoVW vs saliency-based BoVW

For the ADL and IMMED data-sets we present how the different masking approaches influence the results for the Bag of Visual Words framework.

1.6.2.1 IMMED Corpus

The results for the different masking applied on the BoVW framework for this data-sets are depicted in Figures 1.10, 1.11, 1.12, and 1.13. First of all one can notice the effect of manual masking on the performances. The overall performance is about 29.2%. The experimental result of Spatio-temporal-geometric approaches did, as expected, improve in comparison to the baseline method (no mask). It is also important to emphasize that the results obtained from a

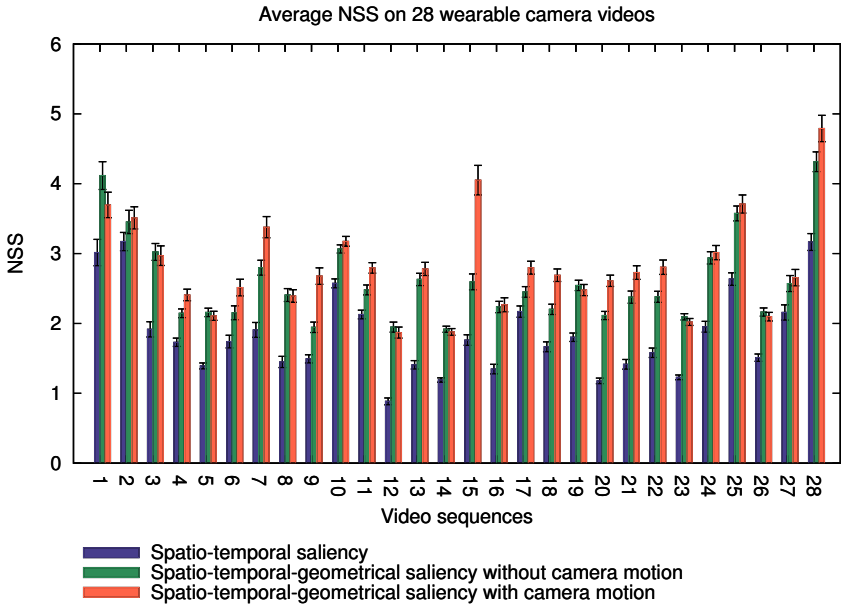
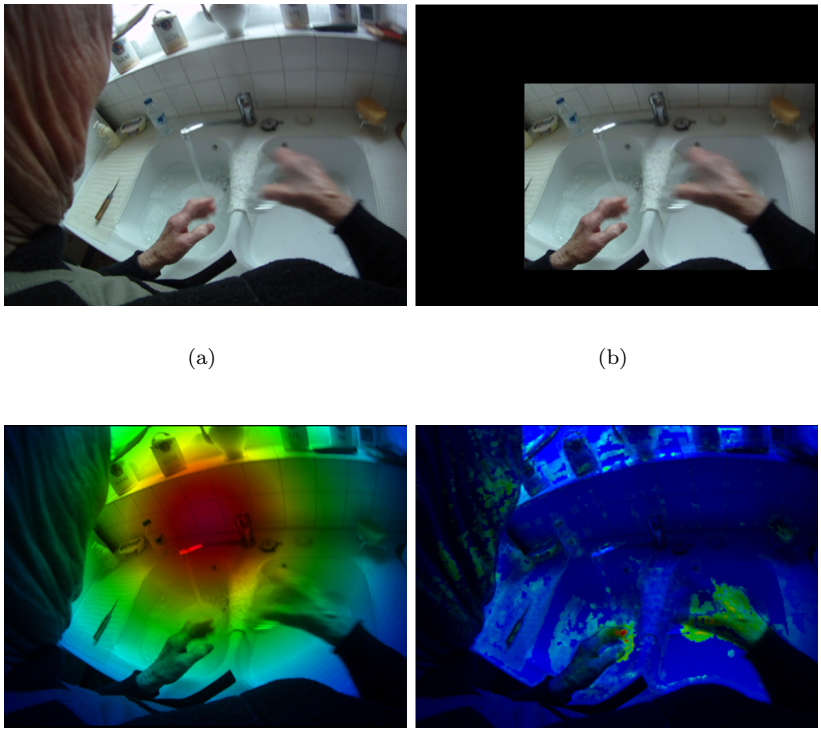


FIGURE 1.9
Objective saliency map correlation with subjective saliency maps.

simple Geometric-based saliency map are better than the one obtained from the analytical approach. We explain this phenomenon by pointing out that, similarly to the Hollywood2 benchmark, the area of interest of the images extracted from the IMMED corpus have been designed to be at the center of the scenes.

1.6.2.2 ADL Data-set

The results for the different masking applied on the BoVW framework for the ADL data-set are depicted in Figures 1.14, 1.15, and 1.16. Similar to the IMMED data-set, the overall performance obtained by manual masking are the best. However the overall experimental results of the Spatio-temporal-geometric approach improved in comparison to the baseline method (no mask).



(a) (b)

(c) (d)

FIGURE 1.10
Visual representation of the 4 different masking approaches for an image from the IMMED corpus. (a) Original frame. (b) Manual masking. (c) Geometric saliency map. (d) Spatio-temporal-geometric saliency map

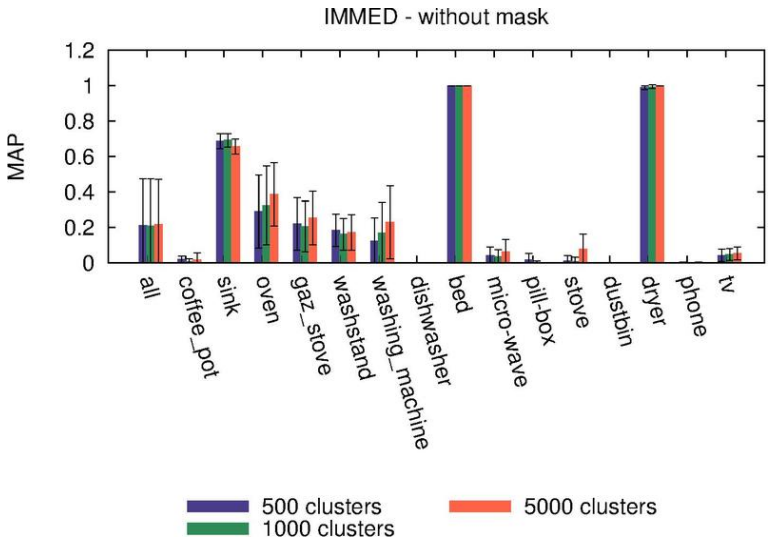


FIGURE 1.11
BoVW results on IMMED database without mask

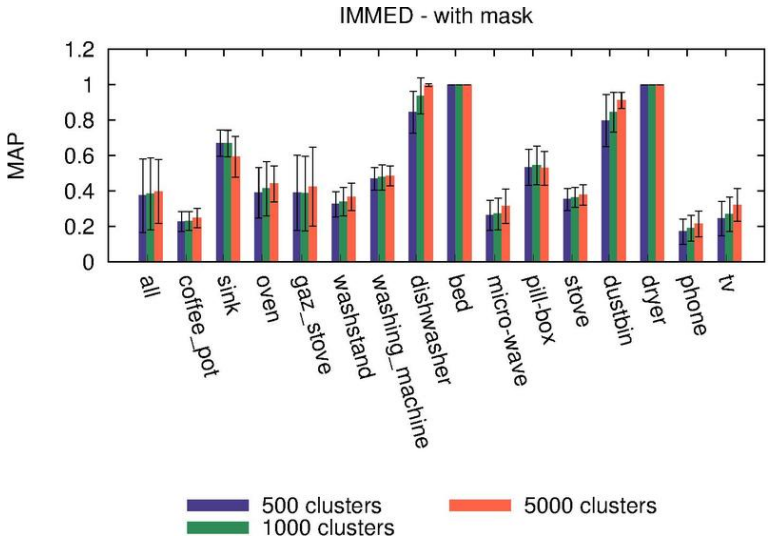


FIGURE 1.12
BoVW results on IMMED database with the ideal mask

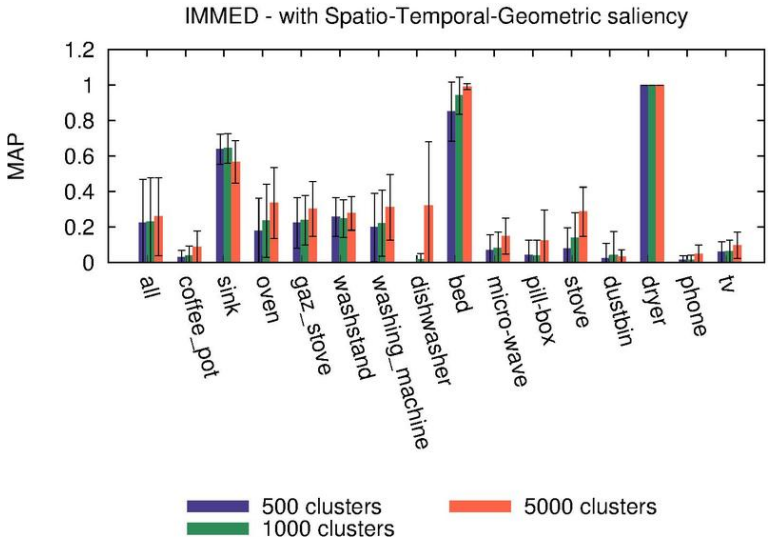


FIGURE 1.13
BoVW results on ADL database with the spatio-temporal-geometric saliency map

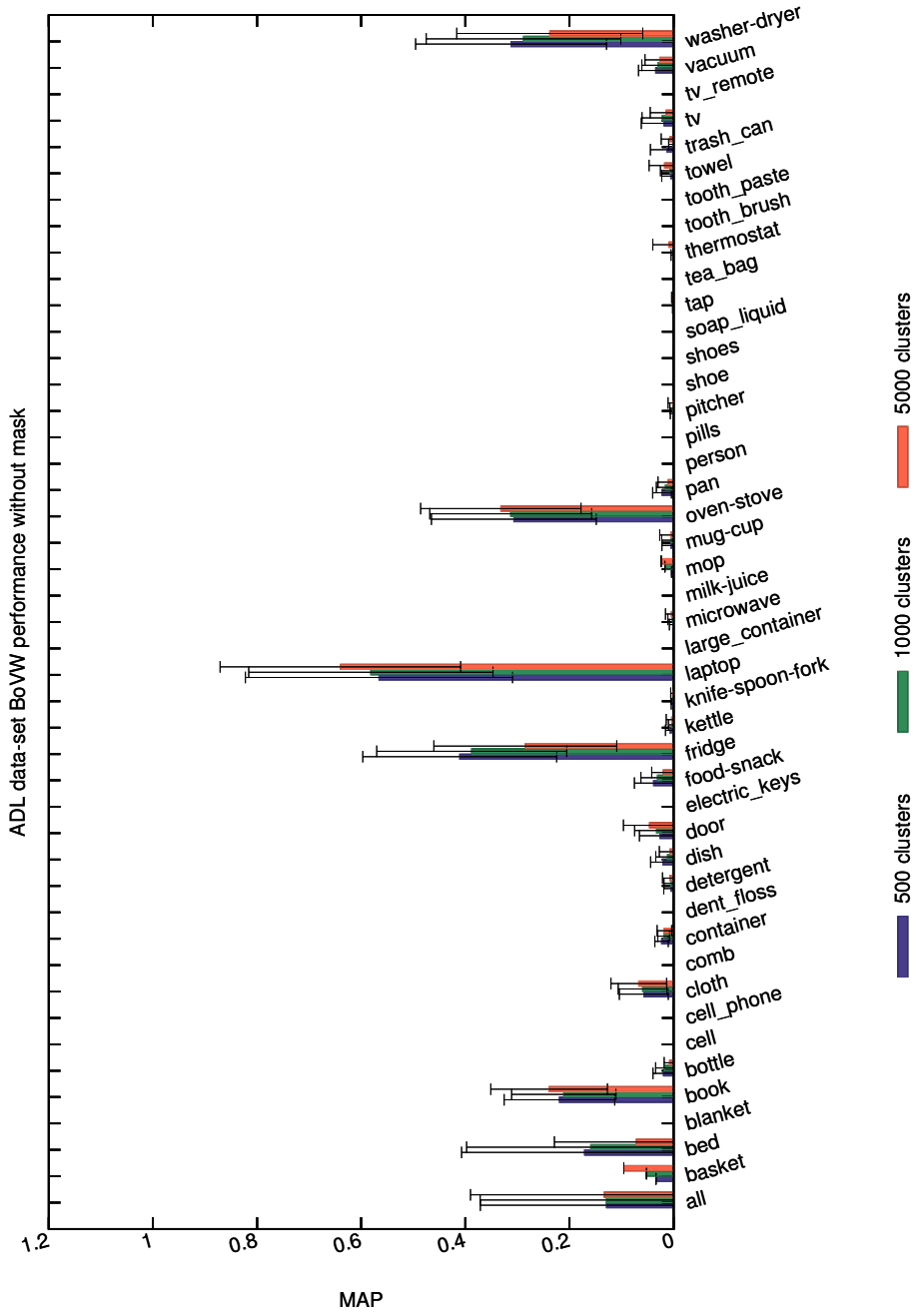


FIGURE 1.14
BoVW results on ADL database without mask

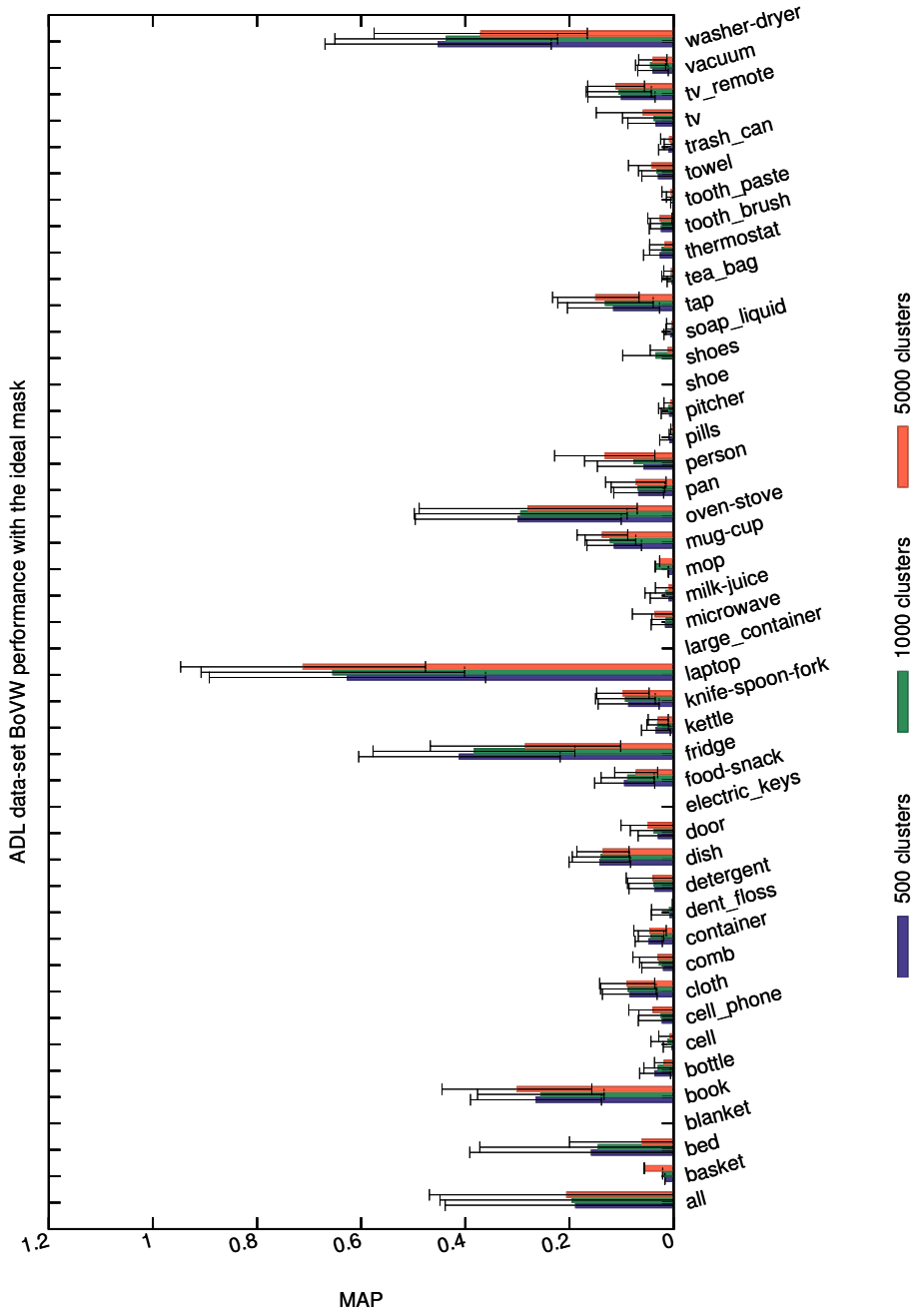


FIGURE 1.15
BoVW results on ADL database with the ideal mask

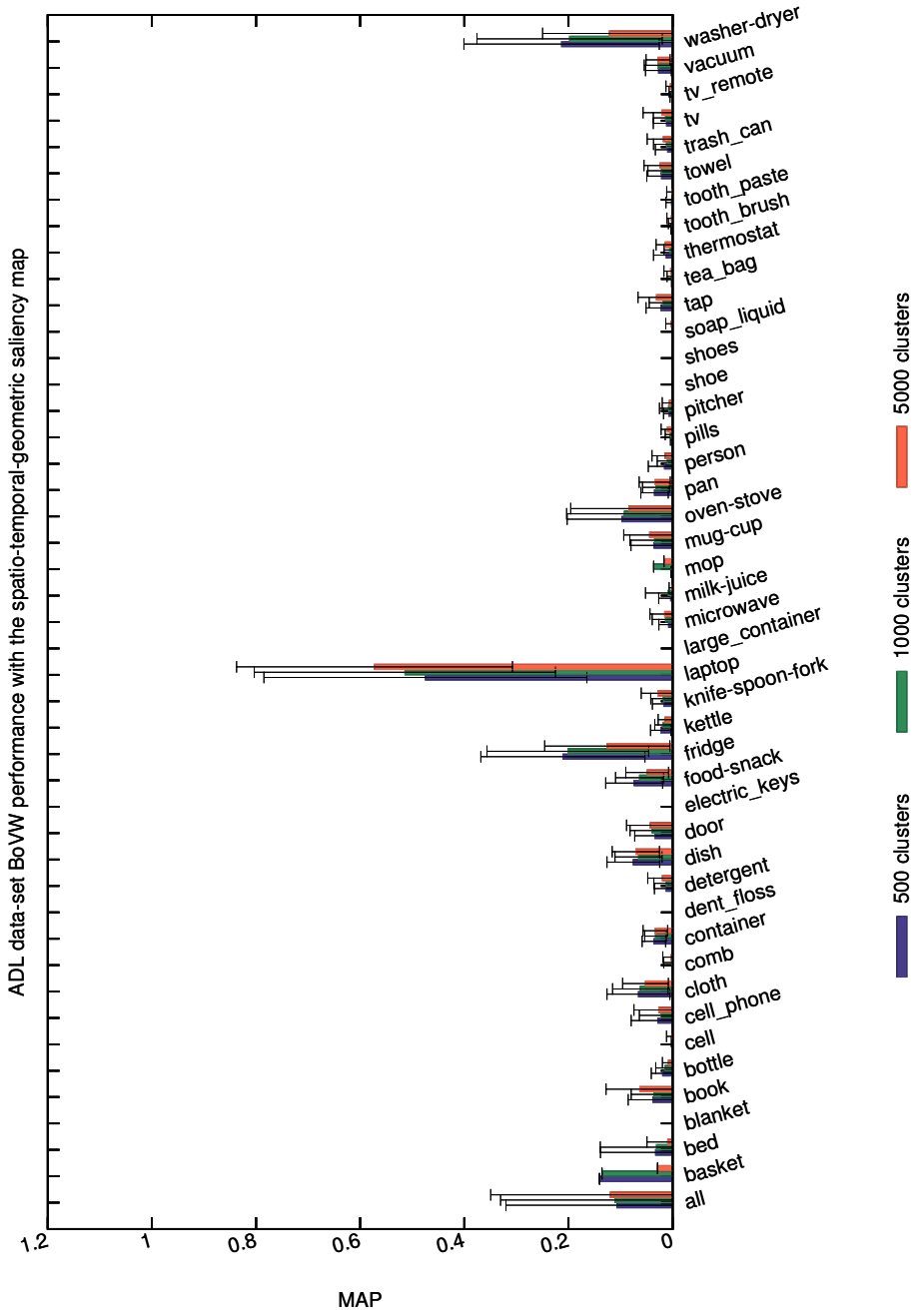


FIGURE 1.16
BoVW results on ADL database with the spatio-temporal-geometric saliency map

1.7 Conclusion

In this work we proposed a saliency based psycho-visual weighting of the BoVW for object recognition. This approach has been designed to identify objects related to IADL on videos recorded by a wearable camera. These recording give an egocentric point-of-view on the upcoming action. This point-of-view is also characterized by a complex visual scene with several objects on the frame plan.

The human visual system functions is a way to process only the relevant data by considering areas of interest. Based on this idea, we propose a new approach by introducing saliency models to discard irrelevant information in the video frames. Therefore we apply a visual saliency model to weight the image signature within the BoVW framework. Visual saliency is well suited for catching spatio-temporal information related to the observer's attention on the video frame. We also proposed an additional geometric saliency cue that models the anticipation phenomenon observed on subjects watching video content from the wearable camera. The findings show that discarding irrelevant features gives better performances when compared to the baseline method which consider the whole set of features in the images.

Thanks to these encouraging results, we believe that our propositions introduce a promising paradigm that can be used in future works to improve the quality of object recognition in complex user-generated videos.

Bibliography

[1] R Albatal, P Mulhem, and Y Chiaramella. Visual phrases for automatic images annotation. In *International Workshop on Content-Based Multimedia Indexing, 2010*, pages 1–6. IEEE, 2010.

[2] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[3] M.Z. Aziz and B. Mertsching. Fast and robust generation of feature maps for region-based visual attention. *IEEE Transactions on Image Processing*, 17(5):633 –644, 2008.

[4] Nicolas Ballas, Benjamin Labbé, Aymen Shabou, Hervé Le Borgne, Philippe Gosselin, Miriam Redi, Bernard Merialdo, Hervé Jégou, Jonathan Delhumeau, Rémi Vieux, Boris Mansencal, Jenny Benois-Pineau, Stéphane Ayache, Abdelkader Hamadi, Bahjat Safadi, Franck Thollard, Nadia Derbas, Georges Quenot, Hervé Bredin, Matthieu Cord, Boyang Gao, Chao Zhu, Yuxing Tang, Emmanuel Dellandrea, Charles-Edmond Bichot, Liming Chen, Alexandre Benoit, Patrick Lambert, Tiberius Strat, Joseph Razik, Sébastien Paris, Hervé Glotin, Tran Ngoc Trung, Dijana Petrovska-Delacrétaz, Gérard Chollet, Andrei Stoian, and Michel Crucianu. IRIM at TRECVID 2012: Semantic Indexing and Instance Search. In *Proceedings of the workshop on TREC Video Retrieval Evaluation (TRECVID)*, page 12p., Gaithersburg, MD, États-Unis, November 2012. CNRS, RENATER, several Universities, other funding bodies (see https://www.grid5000.fr).

[5] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[6] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257 –267, mar 2001.

[7] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbour based image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[8] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints), 2012.

[9] J.-Y. Bouguet. Pyramidal implementation of the lucas kanade feature tracker. *Intel Corporation, Microprocessor Research Labs*, 2000.

[10] 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.

[11] Hugo Boujut, Jenny Benois-Pineau, and Remi Megret. Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion. In Andrea Fusiello, Vittorio Murino, and Rita Cucchiara, editors, *Computer Vision ECCV 2012. Workshops and Demonstrations*, volume 7585 of *Lecture Notes in Computer Science*, pages 436–445. Springer Berlin Heidelberg, 2012.

[12] Y.L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[13] 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.

[14] Alexandre Bur and Heinz Hügli. Optimal cue combination for saliency computation: A comparison with human vision. In *Proceedings of the 2nd international work-conference on Nature Inspired Problem-Solving Methods in Knowledge Engineering: Interplay Between Natural and Artificial Computation, Part II*, IWINAC '07, pages 109–118, Berlin, Heidelberg, 2007. Springer-Verlag.

[15] 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.

[16] Bertrand Delezoide, Frédéric Precioso, Philippe Gosselin, Miriam Redi, Bernard Merialdo, Lionel Granjon, Denis Pellerin, Michèle Rombaut, Hervé Jégou, Remi Vieux, Boris Mansencal, Jenny Benois-Pineau, Stéphane Ayache, Bahjat Safadi, Franck Thollard, Georges Quénot, Hervé Bredin, Matthieu Cord, Alexandre Benoit, Patrick Lambert, Tiberius Strat, Joseph Razik, Sébastien Paris, and Hervé Glotin. IRIM at TRECVID 2011: Semantic Indexing and Instance Search. In *TRECVID 2011 - TREC Video Retrieval Evaluation Online*, Gaithersburg, MD, États-Unis, November 2011. 12 pages - TRECVID workshop notebook papers/slides available at http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html GDR ISIS.

[17] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth. Variability of eye movements when viewing dynamic natural scenes. *Journal of vision*, 10(10), 2010.

[18] Alireza Fathi, Yin Li, and JamesM. Rehg. Learning to recognize daily actions using gaze. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision ECCV 2012*, volume 7572 of *Lecture Notes in Computer Science*, pages 314–327. Springer Berlin Heidelberg, 2012.

[19] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd edition, 2001.

[20] K Grauman and T Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *IEEE International Conference on Computer Vision*, 2005.

[21] D. C. Hood and M. A. Finkelstein. Sensitivity to light. In K. R. Boff, L. Kaufman, and J. P. Thomas, editors, *Handbook of perception and human performance, Volume 1: Sensory processes and perception*, chapter 5, pages 5–1–5–66. John Wiley & Sons, New York, NY, 1986.

[22] Bogdan Ionescu, Constantin Vertan, Patrick Lambert, and Alexandre Benoit. A color-action perceptual approach to the classification of animated movies. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 10:1–10:8, New York, NY, USA, 2011. ACM.

[23] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123, 2005.

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

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

[26] F. Jing, M. Li, H.J. Zhang, and B. Zhang. An effective region-based image retrieval framework. In *ACM International conference on Multimedia*, 2002.

[27] S. Karaman, J. Benois-Pineau, R. Mégret, and A. Bugeau. Multi-layer local graph words for object recognition. In *Advances in Multimedia Modeling*, 2012.

[28] P. Kraemer, J. Benois-Pineau, and J.-P. Domenger. Scene Similarity Measure for Video Content Segmentation in the Framework of Rough

Indexing Paradigm. In *2nd International Workshop on Adaptive Multimedia Retrieval*, Espagne, 2004.

[29] 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.

[30] I. Laptev. On space-time interest points. *International Journal on Computer Vision*, 2:107–123, 2005.

[31] S Lazebnik, C Schmid, and J Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178. Ieee, 2006.

[32] O. Le Meur, P. Le Callet, and D. Barba. Predicting visual fixations on video based on low-level video features. *Vision Research*, 47(19):1057–1092, 2007.

[33] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[34] F. Long, H. Zhang, and D.D. Feng. Fundamentals of content-based image retrieval. In *Multimedia Information Retrieval and Management*, 2003.

[35] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision*, 2(60):91–110, 2004.

[36] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[37] B.S. Manjunath, J.R. Ohm, V.V. Vasudevan, and A. Yamada. Colour and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703715, 2001.

[38] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué. Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, 82(3):231–243, 2009.

[39] F. Mokhtarian and R. Suomela. Robust image corner detection through curvature scale space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1376–1381, 1998.

[40] D Nister and H Stewenius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[41] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[42] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2847 –2854, june 2012.

[43] Marc Pomplun, Helge Ritter, and Boris Velichkovsky. Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception*, 25:931–948, 1995.

[44] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *Computer Vision and Pattern Recognition Workshop*, 2009.

[45] H. Sahbi, J.Y. Audibert, J. Rabarisoa, and R. Keriven. Robust matching and recognition using context-dependent kernels. In *ACM International Conference on Machine Learning*, 2008.

[46] C. Schuldt, I. Laptev, and Caputo B. Recognizing human actions: A local svm approach. In *International Conference on Pattern Recognition*, pages 32–36, 2004.

[47] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, 2003.

[48] T. Starner, B.Schiele, and A. Pentland. Visual contextual awareness in wearable computing. In *ISWC*, 1998.

[49] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[50] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):1–17, 2007.

[51] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, January 1980.

[52] J. van Gemert, C. Veenman, A. Smeulders, and J-M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1271–1283, 2010.

[53] E. Vig, M. Dorr, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *European conference on Computer Vision*, 2012.

[54] D. Wooding. Eye movements of large populations: Ii. deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods*, 34:518–528, 2002.

[55] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.