Gaze based weakly supervised localization for image classification : application to visual recognition in a food dataset

Xin Wang

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Informatique
École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Xin WANG

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Gaze-Based Weakly Supervised Localization for Image Classification: Application to Visual Recognition in a Food Dataset

soutenance prévu le 29 septembre 2017
devant le jury composé de :

M. Patrick Le Callet	Université de Nantes/Polytech Nantes	Rapporteur
M. Philippe-Henri Gosselin	Université de Cergy-Pontoise/ENSEA	Rapporteur
Mme Catherine Achard	Université Pierre et Marie Curie	Examinatrice
M. Chaohui Wang	Université Paris-Est Marne-la-Vallée	Examinateur
M. Fréderic Precioso	Université Nice Sophia Antipolis	Examinateur
M. Nicolas Thome	Conservatoire National des Arts et Métiers	Directeur
M. Matthieu Cord	Université Pierre et Marie Curie	Co-Directeur

PUBLICATIONS

The following publications are included in parts or in an extended version in this thesis:

- Xin Wang, Nicolas Thome, and Matthieu Cord (2017). "Gaze Latent Support Vector Machine for Image Classification Improved by Weakly Supervised Region Selection." In: *Pattern Recognition*, pp. –. DOI: https://doi.org/10.1016/j.patcog.2017.07.001. URL: http://www.sciencedirect.com/science/article/pii/S0031320317302625.
- Xin Wang, Nicolas Thome, and Matthieu Cord (2016). "Gaze latent support vector machine for image classification." In: *IEEE International Conference on Image Processing (ICIP)*, pp. 236–240.
- Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso (2015a). "Recipe recognition with large multimodal food dataset." In: *IEEE International Conference on Multimedia & Expo Workshops*, pp. 1–6.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 GENERAL CONTEXT

Eyes are the windows to the soul. For the artificial intelligence agent, the eye and soul are based on the computer vision technology. Computer vision has achieved quite a lot of advancements in many industrial areas, including but not limited to self-piloting automobile, medical image analysis, security monitoring, etc. Generally, computer vision is about perceiving the visual environment and understanding the visual content. Researchers generalize the complex practical problems of computer vision into several basic tasks, including category recognition, object localization, object detection, object segmentation, etc. As the fundamental understanding of the visual content, image classification has been deeply researched and a great number of applications have been built on it. The goal of image classification is to predict what is the semantic category of an image according to its visual content. In this thesis, we focus on the problem of image classification.

The huge amount of stock and increment visual information on the Internet, typically known as *big data*, makes this problem extremely challenging. In the white paper of Visual Networking Index (Cisco 2016), Cisco shows a statistic that video content holds around 70 percent of the Internet traffic by the year of 2015. Based on the status quo, they estimate that every second, nearly a million minutes of video content will cross the network by 2020. These big data derive from the soaring increase of portable devices such as mobile phones, digital cameras and Internet services such as social networks, video games, etc. For instance, 350 million photos are uploaded per day to Facebook[1] and 80 million to Instagram[2] . Mining the value from such tremendous data could become possible, only if we can design

1 http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9?IR=T
2 https://maximizesocialbusiness.com/definitive-instagram-statistics-23286/

proper algorithms for handling such a high order of magnitude of data. Recently, due to the rapid development of deep learning, statistical machine learning methods have achieved amazing results on this issue, even surpassed human-level performance (He et al. 2015a). This excellent result makes the machine learning methods began to get large-scale deployment and application in the real industrial products. It also attracts the world's best technology companies, *e.g.* Google, Facebook, Microsoft, Amazon, Baidu, etc., as well as a large number of startups, research projects, public infrastructure services, to promote or benefit from the development of statistical machine learning.

1.2 VISUAL RECOGNITION: SHALLOW, DEEP AND WEAKLY SUPERVISED LEARNING

Image classification is a challenging task for machines. For humans, this problem is a natural ability, but the current science has not yet fully understood the human eye - brain coordination mechanism, and therefore can not be copied to the machine. This problem is also complicated for ordinary logic-based computer programs because the representation of an image or object may vary greatly (such as image rotation, brightness variation, background noise, object deformation, etc.), but the corresponding semantics may be the same. The main challenge is that low-level image representations (i.e. the pixels) are not discriminative enough to directly predict semantic-level concepts, generally known as *semantic gap* (Smeulders et al. 2000). At present, the state-of-the-art solution is based on the machine learning model. The machine learning model describes the original picture as a mapping from a string of numbers (pixel representation of the image) to the image semantic label. Then the most important step is to learn the mapping function $f : x \rightarrow y$ from these data. Note that this step does not require explicit programming. After the learning is completed, when the machine sees a new image, it uses the mapping function to interpret x as the semantics of y, for example in the Fig. 1.1 the whole image content is x, and y is composed of *Eiffel tower*, *person*, *tree* etc. This is the basic process of machine learning in image recognition.

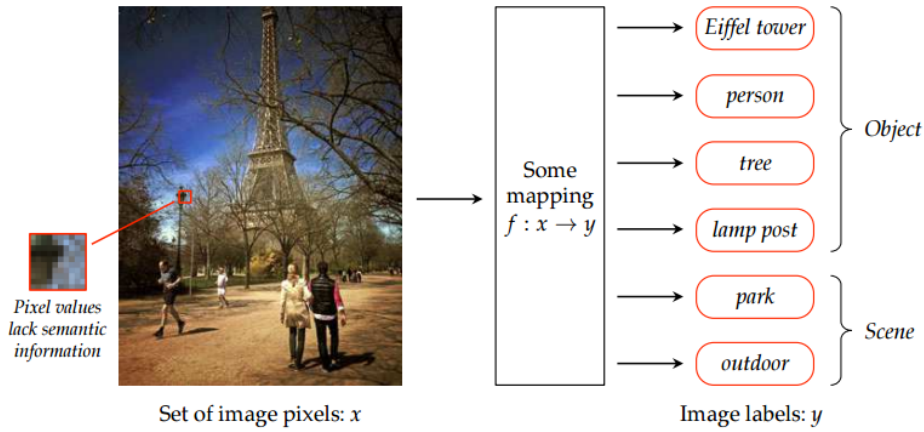


Figure 1.1: The image annotation problem. The challenge of image annotation is to find a mapping that bridges the semantic gap between raw image pixels and semantic concepts, such as objects and scene categories. (Credit Hanlin Goh)

One of the key part in this model is defining the image description, namely the input feature x to the machine learning model. A good representation makes the classification become easy. Before the year of 2012, most of the image classification features were hand-crafted. The intuition of these features often originated from the human observation and trials, calculating some fixed quantitative indicators according to the image pixel value, *e.g.* the most popular one: Scale-invariant feature transform (SIFT) (Lowe 2004). Then, considering the computation efficiency, people often constituted a compressed vector to describe the image. The most popular approach was the Bag of Words (BoW). It was first applied in the text retrieval task (Salton and McGill 1986). The basic idea is to express the text as a histogram vector of the frequency of the words. To extend this idea to image classification, a key issue is the definition of visual vocabulary. A famous solution is to cluster the features in the image to obtain a man-made visual vocabulary dictionary, then representing an image as a histogram vector (Ma and Manjunath 1999; Sivic and Zisserman 2003).

Deep learning has achieved great success in the era of big data. The dominance of deep models is witnessed in the fields of face recognition (Learned-Miller et al. 2016), machine translation (Zhou et al. 2016), speech recognition (Saon et al. 2016), and even the Go game (Silver David, Huang Aja, and et.al 2016). One typical example is the success of deep convolutional neural network (DCNN) in computer vision. Since the year of 2012, DCNN-based image descriptions largely outperforms hand-crafted features. The first breakthrough came from a research team in the Toronto university (Krizhevsky, Sutskever, and Hinton 2012), who achieved 15% accuracy on ImageNet Large Scale Visual Recognition Competition (ILSVRC 2012), outperforming the second-place BoW-based method by ten percentage points. From the AlexNet (Krizhevsky, Sutskever, and Hinton 2012) to the state-of-the-art deep Residual Networks (He et al. 2016), the DCNN has much outperformed the traditional hand-crafted feature-based machine learning methods on the largest classification competition ImageNet Large Scale Visual Recognition Competition (ILSVRC) (Deng et al. 2009). It is worth noting that ImageNet is a super large scale dataset contains more than 10 millions of labeled images. To today (2016 ILSVRC), deep convolution neural network has reached a 3% error rate, even significantly exceeding the human-level error rate of 5.1%. Moreover, deep models trained on ImageNet can also be applied effectively to different target domain or different tasks by *transfer learning* (Yosinski et al. 2014). As a result, state-of-the-art results on standard benchmarks are nowadays obtained with deep features as input. Recent studies show that *fine-tuning* and *data-augmentation* can further boost the performance of the transferred models (Chatfield et al. 2014).

In fact, DCNN is not a new technology that suddenly appears. As early as 1957, Rosenblatt presented the Perceptron to simulate brain neurons (Rosenblatt 1957). Later in 1980, Fukushima presented the paper of Neocognitron (Fukushima 1980), using the concept of receptive field on the basis of Perceptron to construct the prototype of DCNN (Hubel and Wiesel 1962). In 1989, Lecun used DCNN to develop a handwritten digital identification system for bank automatic identification checks(LeCun

et al. 1989). The outbreak of DCNN in the 2010s can be attributed to two important factors:

1. A large number of labeled data can be trained with large deep neural networks without over-fitting.

2. The rise of new computing devices, such as GPUs, significantly reduces the time required to train large neural networks.

A DCNN is composed of several non-linear transformation layers to convert an input image to the target value, *e.g.* image label. The neural network contains a large number of parameters, but these parameters are all learned from the data, so that the feature extraction and the final task is a more reasonable joint training process, which is often referred as the *end-to-end training*, rather than the traditional method of separate training. The hierarchical representations progressively abstract image features, making the original highly linear inseparable low-level features become approximate linear separable. This is why we observe that applying a linear classification model on the representations of the neural network at the deeper layer can also achieve a good classification result.

Despite the great success of DCNN, one of its bottlenecks is the lack of spatial invariance. Spatial invariance refers to the ability for dealing with non-central, scale-variant, clutter objects in the image dataset. In the Fig. 1.2, we see that the objects in the ImageNet are centered and large with respect to the whole image, but not for the images in the natural image datasets such as PASCAL VOC (Everingham et al. 2015) and MS COCO (Lin et al. 2014). In order to improve the accuracy of classification, an intuitive idea is to select the image area associated with the target semantics to classify. Because clutter information decreases the discriminative power of the model. In this case, expensive annotations such as bounding boxes are often used to localize the target object. Clutter information is subsequently filtered out by omitting the information outside the bounding boxes. This idea is first presented in this paper of Russakovsky et al. (Russakovsky et al. 2012a): extracting the feature in the bounding box area of the target object and finding a significant improvement over the classification accuracy of the extracted feature from the whole image. This illustrates the

Figure 1.2: Examples of *car* images from ImageNet, PASCAL VOC 2012 and MS COCO.

consistency between the region and the target object ensures the quality of the extracted features, and the local invariance of the object is the factor that the image classification system needs to consider. For the DCNN, since most models are pre-trained on the ImageNet, a direct transferring application is not compatible with the non-centered dataset. (Oquab et al. 2014) exploits the bounding box supervision to train object-centric deep classifiers, which perform better on the PASCAL VOC. Recently, attempts have been made to overcome this limitation by encoding local information by following the design of Bag-of-words (BoW): (He et al. 2015b; Gong et al. 2014) proposed BoW models with deep features as local region activations and (Arandjelovic et al. 2016) developed BoW layers.

A potential limitation for promoting the local invariance using full annotation is that it is extremely time-consuming and scarce. In the Fig. 1.3, we compare the annotation time and quantity of the common data annotation. We observe that the annotation with rich information, *e.g.* bounding box or pixel-wise segmentation, is time-consuming and scarce, while the annotations with coarse information, *e.g.* noisy label or image-level label, is time-saving and abundant. The ideal solution is to use coarse annotations to get comparable results as using rich annotation. The method for achiev-

ing this goal is named as Weakly Supervised Learning (WSL). In this thesis, we are interested in a specific problem of image classification improved by weakly supervised localization, which is still in the form of bounding box to locate the object but no longer needs human manual labeling.

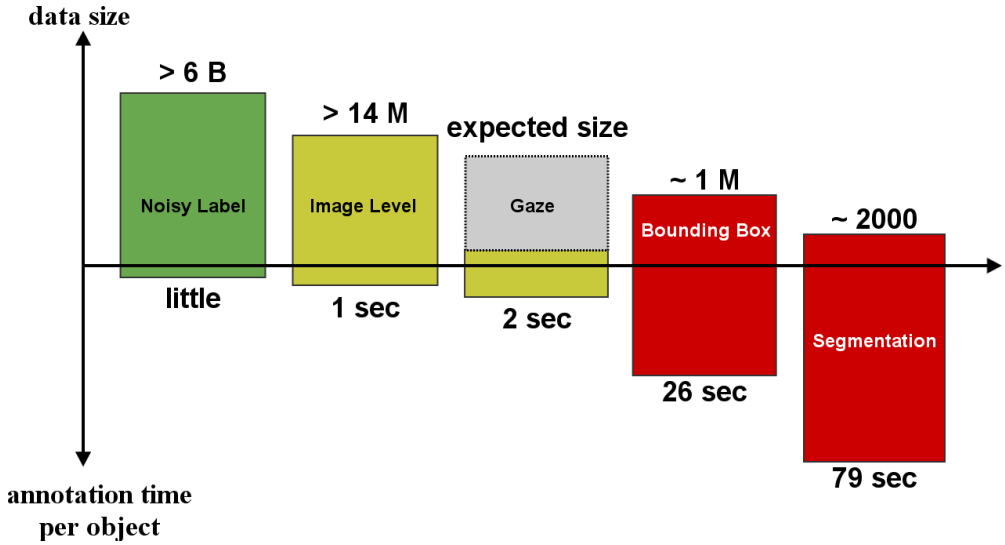


Figure 1.3: Time cost and data size of different kinds of image annotations. (Figure data is compiled from (Russakovsky et al. 2016; Papadopoulos et al. 2014; Matthew Blaschko, Pawan Kumar, Ben Taskar 2013))

1.3 GAZE ANNOTATION AND WEAKLY SUPERVISED LOCALIZATION

Selecting relevant regions from images with only image label is a challenging task for training a WSL model. Noise label is fine at both labeling time and quantity, but the information it carries may not be suitable for training a robust model. Image label is commonly used in WSL models, but it carries no localization information except for categorical information. Comparing to image label, the gaze carries object localization information (Yun et al. 2013) and can be easily extended to carry image label as well (Papadopoulos et al. 2014). Also, as shown in Fig. 1.3, comparing to

commonly seen annotations, labeling gaze annotation costs little time as image label. Although the quantity of gaze annotation is not as much as image label, we anticipate that gaze annotation quantity will increase in the future by considering its advantages. Researchers have already begun to think of leveraging human visual pattern for improving computer vision performance. One possible choice is to analyze the eye movement and track where exactly human watch using an eye-tracker. An eye-tracker is a device that incorporates illumination, sensors and processing to track eye movements and gaze point. Recently, the use of near-infrared light allows for accurate, continuous tracking regardless of surrounding light conditions. In this thesis, we use a small cuboid-like eye-tracker, which can be placed just under the screen of a laptop without any intrusion towards the human eyes. In a nutshell, human eyes are illuminated by a light source then reflect the light, a camera then captures an image of the eye showing these reflections. By identifying the reflection on the cornea and in the pupil, and combining with other geometrical features of reflections, eye-tracker is able to calculate the gaze direction. In this thesis, we consider gaze features recorded by an eye-tracker device, which present two useful properties: one is that gaze features, when collected from people asked to identify a semantic category in an image, contain useful information about the position of the target objects or relevant regions for classification. We then focus on image classification improved by weakly supervised gaze-biased region selection.

1.4 CONTEXT OF THIS THESIS

This thesis is registered under the ANR project VIsual Seek for Interactive Image Retrieval (VISIIR). This project aims at exploring new methods for semantic image annotation, especially *gaze*. This project includes several research topics: unsupervised bio-inspired image representations, visual salience, eye-driven machine learning system, multimodal food recipe retrieve. Our thesis covers all topics, but mainly study the latter two issues.

In this dissertation, our classification algorithms are all validated on a *multimodal* recipe dataset. Here the *multimodal* means that the recipe is

represented by both the visual image and the textual recipe. Food categorization is an emerging topic in the multimedia research community. The mainstream methods for food classification are based on image visual classification (Fig. 1.4). A classical application scenario of automatic food image categorization is to answer the question in the restaurant: "What is this dish?" Although the food image categorization is in spirit the same as image classification, food categorization remains a difficult problem because of the diversity of textures, large variation of shape, complicated mixture of elements, etc. Intuitively, as a complement information of images, multi-modal data such as ingredients, recipe text, restaurant geolocalization are exploited to build more robust classification systems (Jingjing Chen 2016; Min et al. 2016). Our team builds this large scale multimodal dataset and fuse multimodal information to get a deep understanding of this distinctive recipe dataset. On the other hand, we used an eye-tracker to annotate part of the images in this dataset for validating the effectiveness of our gaze-based weakly supervised classification system.

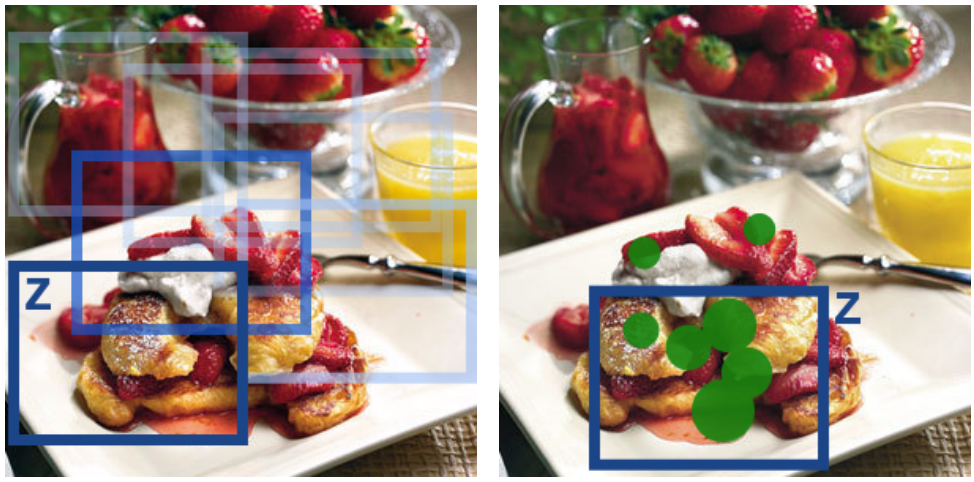
(a) *tiramisu* (b) *tiramisu* (c) *pancakes*

Figure 1.4: Sample of food images. The objective of image categorization is to answer the right food category name.

This thesis contains contributions in three areas: multimodal food dataset recognition based on deep learning, weakly supervised learning and gaze-based weakly supervised localization.

Deep learning has much outperformed the traditional hand-crafted feature-based machine learning methods in various computer vision problems. For training a deep model from scratch, it requires a large number of data to achieve a reasonable result. Fortunately, transfer learning and fine-tuning can help us to train with reasonable scale image datasets. Deep learning has also achieved the state-of-the-art performance on various natural language processing tasks. The great performance of deep learning makes it attractive to process multimodal data. For this reason, we first collect a large scale multimodal food dataset, called UPMC Food-101, including pairs of visual image and textual recipe. Based on this dataset, we perform both shallow and deep learning with visual and/or text information. We learn from the result that the deep-based methods outperform the shallow methods in classification and retrieve tasks with a large margin.

The weakly supervised learning is a framework where the model learns to capture aspects of the data that are not labeled in the training data. Learning from weakly labeled data covers several practical aspects towards the development of powerful learning machines. It helps to reduce the amount of annotated information used for learning and is promising for making full use of the data. Handling weakly labeled data generally requires learning a model with latent variables to model hidden factors for compensating the weak supervision. For image classification with weak localization scheme, the hidden factor is the possible localization of the target object. Fig. 1.5a illustrates the intuition: the sub-region which contains the object is treated as a latent variable. Based upon the regional image feature, a model learns to find the region Z, which is the most consistent with the image semantic. This procedure only requires the global image label, but outputs simultaneously the classification result and the inferred region. Inspired by the human recognition process, integrating the gaze into the weakly supervised localization scheme for image classification



(a) Weakly supervised localization (b) Weakly supervised localization with gaze

Figure 1.5: Given an image with image-level label (Fig. 1.5a) or gaze annotation (Fig. 1.5b) for training an image classifier with weakly supervised localization scheme. In Fig. 1.5a, the model searches for the semantic region Z based on regional image feature, while in Fig. 1.5b, the model tends to localize a region Z where the image feature and the density of gaze are balanced. (Notation: a rectangle represents a candidate region of an object, and the darker the color is, the higher the possibility is. The region Z is the most possible region. The green circle represents the gaze, whose radius reflects the duration.)

is promising. Compared with the global image label, the gaze annotation costs approximately the same amount of time to acquire, but has a weak localization information that an image label does not have. In this thesis, we propose using gaze-based weakly supervised localization for image classification. The idea is shown in Fig. 1.5b: when training a classifier, the model tends to localize a region where the image feature and the density of gaze are balanced. Comparing to weakly supervised localization with only image label, we find that our proposed method can learn to localize more semantic meaningful object regions. Along with the weakly supervised localization strategy, the image classification performance also increases

significantly. These results are verified on several benchmark datasets, including a subset of UPMC-Food101 annotated with an eye-tracker, which is called UPMC-G20.

The outline of this thesis is as follows:

- In Chapter 3, we present our multi-modal web-based food classifier and the large scale food dataset UPMC Food-101. We first propose a large scale food-related multimodal dataset: each instance in this dataset contains an image and a corresponding recipe text. In the multimodal context, we consider using the weakly aligned visual and textual representation to retrieve the recipes. The content of this chapter is based on (Wang et al. 2015a).

- In Chapter 4, we introduce a gaze-based model for performing weakly supervised localization image classification. Regions in the images are modulated by gaze information for indicating the possibility of having a target object in this region. This model needs only gaze annotation for learning, while the test phase is gaze free. This property is useful since we can apply the trained model onto any unseen image without gaze annotation. This chapter is based on (Wang, Thome, and Cord 2016).

- In Chapter 5, we propose to improve the model in Chapter 4 by two factors: 1) localizing the information in the negative training images using gaze, which makes full use of the gaze annotation for all training images, 2) selecting multiple regions instead of a single region for localizing the object, which makes a set of regions benefit the gaze modulation. This model strengthens the weakly supervised region selection capacity and leads to a better generalization performance. For consolidating the robustness of our model, we annotate part of the UPMC Food-101 images with an eye-tracker. We make this dataset because food is also a complicate object made up of various ingredients and often with several common backgrounds. The extended model and the gaze-based dataset UPMC-G20 are published in (Wang, Thome, and Cord 2017).

Before presenting our contributions, we provide relevant background in Chapter 2 for image classification with machine learning techniques, weakly supervised learning (WSL), eye-tracking analysis and the interdisciplinary fields of these disciplines. Finally, the conclusion and perspectives are presented in Chapter 6.

2

RELATED WORKS

Visual understanding is a key component of artificial intelligence and has been researched for a long time. In this chapter, we relate our work with the previous works on the image recognition machine learning techniques, multimodal food data understanding, eye-tracking research and weakly supervised learning framework.

2.1 IMAGE RECOGNITION

2.1.1 *Bag of visual words (BoVW) models*

In the history of solving image recognition problem using statistical paradigm, the first milestone of image recognition derives from the *Bag-of-Visual-Words* (BoVW) model. The original BoVW is used for content based image indexing and retrieval (CBIR) in videos (Sivic and Zisserman 2003), but the design framework has a profound influence in the research of computer vision in the following decade. In the BoVW (shown as Fig.2.1, the images are first represented by pixel-level hand-crafted feature descriptions, *e.g.* Scale-Invariant Feature Transform (SIFT) (Lowe 2004), Histogram of oriented gradients (HoG) (Dalal and Triggs 2005), Speeded up robust features (SURF) (Bay et al. 2008), etc. Then an unsupervised cluster method applies on the features for generating a dictionary and a corresponding visual-words counting-histogram. This histogram can be regarded as a naive version of BoVW. This histogram is further fed into models as input for learning to solve specific tasks. Along with the BoVW features, the learning methods can be adapted to various computer vision applications, including object matching (Lowe 2004), image classification (Csurka et al. 2004), human action recognition (Wang and Mori 2009), facial expression recognition (Fasel, Monay, and Gatica-Perez 2004), medical images (Wang

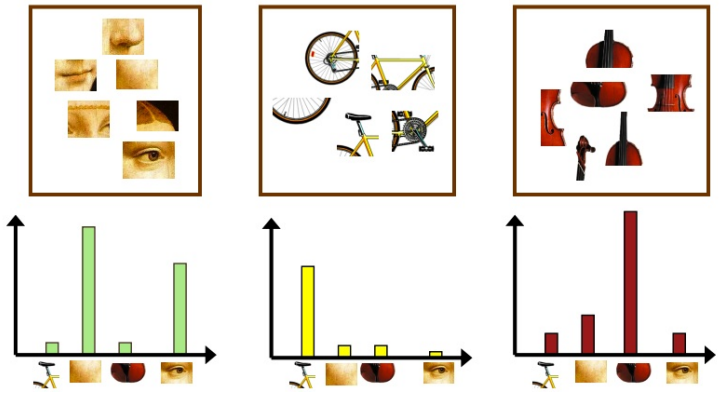


Figure 2.1: Bag of Visual Words (BoVW) scheme. (Credit: cs143 course of Brown University)

et al. 2011), sport image analysis (Kesorn and Poslad 2012), 3D image retrieval and classification (Li and Godil 2010; Toldo, Castellani, and Fusiello 2009), image quality assessment (Ye and Doermann 2012), etc.

Despite the great success of BoVW model, this technique suffers from two intrinsic drawbacks: one originates from the BoVW representation. As the SIFT-like features mainly describes the local low-level characteristics of images, high-level semantic information is not described. The other drawback originates from the learning methods. Building the mapping from low-level statistic to the target semantic is a black-box. The desired hierarchy architecture should learn step-by-step from the low-level pixel to high-level semantic. In fact, the hierarchy architecture achieved great success in almost all research ares of artificial intelligence since 2006, which are collectively referred as the deep learning methods.

2.1.2 *Deep learning in computer vision*

During recent years, *deep learning models* have emerged as another milestone of image recognition. The prototype of deep learning dates back to 1950s, which is known as *Perceptron* (Rosenblatt 1957). For computer vision, deep convolutional neural networks (DCNN) (LeCun et al. 1989) achieve

great success. As shown in Fig. 2.2, a DCNN is generally composed of several non-linear transformation layers to convert an input image to the target value, *e.g.* image label. Among these layers, the most important one is the convolution layer. The convolution layer considers the spatial structure of image data, encoding the content of the image from the local to the whole. Comparing with the Perceptron, CNN is notable for its convolutional layer. Essentially, convolutional layer is able to learn local filters. When we stacked the convolutional layers, the local filter in the deeper layer becomes global filter with respect to the original image. That's also why the power of CNN shows up when the network goes deeper. However, as deep CNN is calculation-intensive, only until the year 2012, when the calculation based on the novel computational devices (*i.e.* GPUs) and the parallel computation techniques is implemented, the deep CNN largely exceeds BoVW-based model for the first time in the world's largest image recognition competition ILSVRC (Krizhevsky, Sutskever, and Hinton 2012). This huge success is one of the most meaningful time nodes for the prosperous deep learning research we are currently witnessing.

Deep learning is popular for following reasons:

1. Powerful generalization ability. Since DCNN won the ImageNet Large Scale Visual Recognition Competition (ILSVRC 2012), the state-of-the-art of most of the computer vision problems are based on DCNN. For example, recently, experimental results show that for image recognition tasks, machine can perform even better than human beings (He et al. 2016). The dominance of deep models is also witnessed in the fields of face recognition (Learned-Miller et al. 2016), machine translation (Zhou et al. 2016), speech recognition (Saon et al. 2016), and even the Go game (Silver David, Huang Aja, and et.al 2016).

2. End-to-end learning. This is an useful property for learning from data without human intervention. This means that the learning starts from the raw data and ends at the target concept, in which all parameters are learnable.

3. High quality transferable representation. Deep models trained on large scale dataset can often be applied effectively to different tar-

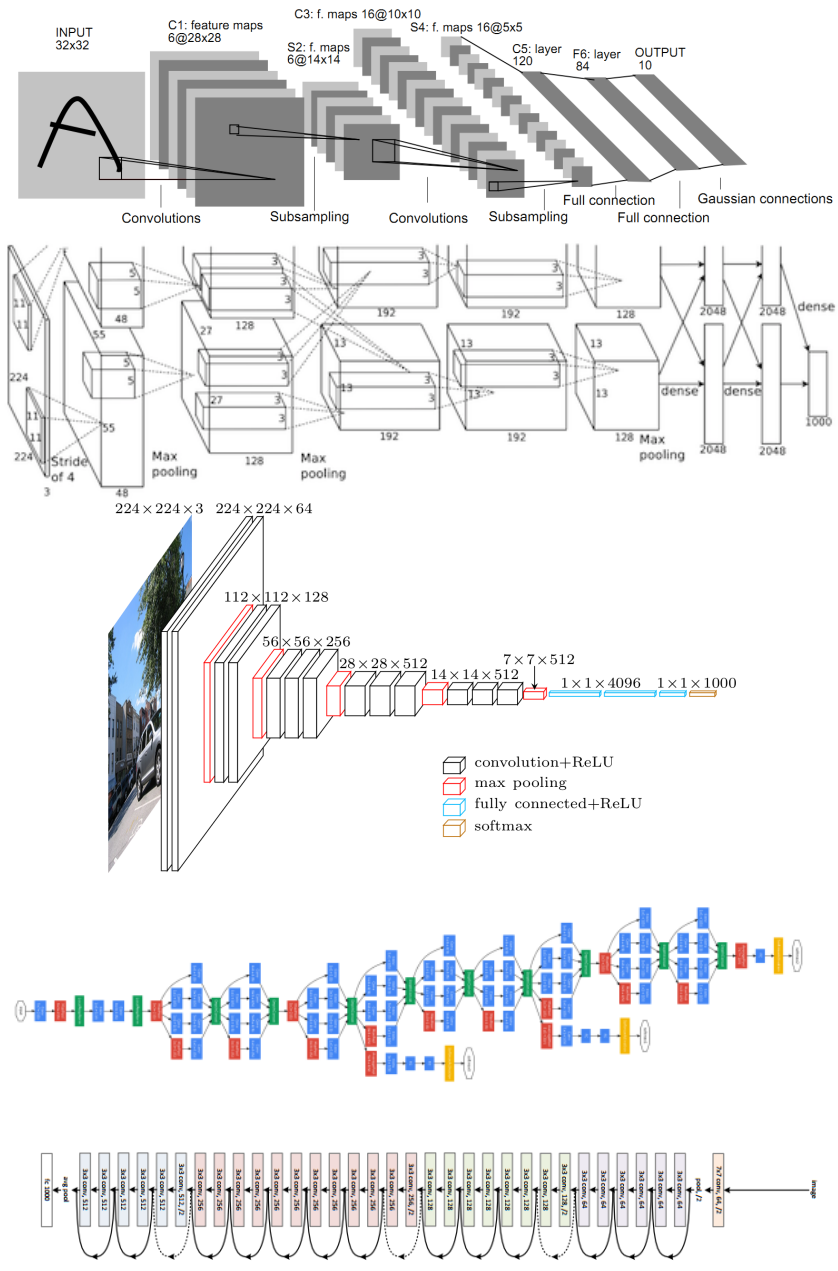


Figure 2.2: A list of classic deep convolutional neural networks. From top to bottom: LeNet-5(LeCun et al. 1989), AlexNet (Krizhevsky, Sutskever, and Hinton 2012), VGG (Simonyan and Zisserman 2015), GoogleLeNet (Szegedy et al. 2015), Residual Neural Network (ResNet) (He et al. 2016).

get domain or different tasks by transfer learning, *e.g. computer vision* (Yosinski et al. 2014). As a result, state-of-the-art results on standard benchmarks are nowadays obtained with deep features as input. Recent studies show that fine-tuning and data-augmentation can further boost the performance of the transferred models (Chatfield et al. 2014).

However, deep learning is "data hungry" because a typical deep neuron network has more than millions of parameters. If the model for a given application is trained with a small-scale dataset, it will probably cause over-fitting. Except for conventional L1-regularization, L2-regularization and data augmentation, several regularization methods designed for the structure of the neural network are proposed to overcome the over-fitting (Srivastava et al. 2014; L.Wan et al. 2013). For making deep learning running on the small-scale dataset, *transfer learning* (Yosinski et al. 2014) strategy is one of the practical strategy. Transfer learning means improving the target predictive function in the target domain using knowledge learned from the source domain. This strategy largely reduces the effort for re-training a network. For example in our work, transferring a model learned on ImageNet, the world's largest labeled image dataset, to UPMC Food-101 outperforms traditional hand-crafted image representations by a large margin. The improvement is reasonable since the deep hierarchical architecture ensures that the models learn from the basic representations to high level representations, leading to a better generalization on other data distribution or other tasks. Furthermore, *domain adaptation* techniques, such as taking a model already trained on the source domain, then fine-tuning part of the parameters by the data of the target domain, can better generalize the model on the target domain than simply taking the model trained on the source domain or directly training on the target domain. In the computer vision community, a lot of experiments claim that taking the models learned on ImageNet leads to substantial progress for other datasets or even other vision tasks (such as segmentation, motion analysis, etc.) with transfer learning or domain adaptation.

2.2 MULTIMODAL FOOD DATA UNDERSTANDING

Our work is in the scope of multimodal food dataset collection, food image categorization methods and food-related applications. In this section, we review some relevant researches on this purpose.

2.2.1 *Multimodal food datasets*

There is an increasing demand of food-related dataset for applications like dietary assessment, computational cooking, recipe retrieval, etc. We summarize current multimodal food datasets in Table 2.1. Most of the datasets are purely image dataset. One of them is the Pittsburgh Food Image Dataset (PFID) (Chen and al. 2009) dataset, containing 4556 images out of 101 fast food categories. Another one is UNICT-FD889 dataset (Farinella and all 2014) that has 3583 images out of 889 distinct dishes. UEC-Food100 (Kawano and Yanai 2014c) contains 100 categories of food images, each category contains about 100 images, mainly Japanese food categories. Expanding UEC-Food100 by a food-image retrieve method results in a new dataset UEC-Food256 (Kawano and Yanai 2014d), which contains more than 100 images out of 256 categories is proposed. ETHZ Food-101 (Bossard and al. 2014) contains 101,000 images out of 101 categories. (He et al. 2014) propose 1453 images with 42 categories. (Myers et al. 2016) proposes a classification dataset Food201-MultiLabel, and a segmentation dataset. Food201-MultiLabel dataset contains nearly 50,000 images out of 201 categories. Food201-Segmented contains nearly 13,000 images out of 201 categories. (Pouladzadeh, Yassine, and Shirmohammadi 2015) contains 3000 images with 23 food categories. (He, Kong, and Tan 2016) proposes 15,262 images of 55 categories.

Recently, multimodal food datasets with richer information are proposed. VIREO Food-172 (Jingjing Chen 2016) contains 110241 images out of 172 categories and 353 ingredient labels. Yummly-28K (Min et al. 2016) contains nearly 28k items (image + recipe name) out of 3000 ingredients (2208 visible + 792 non-visible), 16 kinds of cuisines, 13 kinds of recipe courses. (Luis Herranz 2015; Xu et al. 2015c) proposes a dataset containing a total of 187

	nb. images	nb. categories	ingredient	recipe	geolocalization
(He et al. 2014)	1,453	42	×	×	×
(Pouladzadeh, Yassine, and Shirmohammadi 2015)	3,000	23	×	×	×
UNICT-FD889 (Farinella and all 2014)	3,583	889 dishes	×	×	×
PFID (Chen and al. 2009)	4,556	101	×	×	×
(He, Kong, and Tan 2016)	15,262	55	×	×	×
Food201-Multilabel (Myers et al. 2016)	50,000	201	×	×	×
UEC-Food100 (Kawano and Yanai 2014c)	100,000	100	×	×	×
ETHZ Food-101 (Bossard and al. 2014)	101,000	101	×	×	×
UEC-Food256 (Kawano and Yanai 2014d)	256,000	256	×	×	×
VIREO Food-172 (Jingjing Chen 2016)	110,241	172	√	×	×
Yummly-28K (Min et al. 2016)	28,000	16 cuisines/13 courses	√	×	×
(Luis Herranz 2015; Xu et al. 2015c)	∼ 20,000	701 dishes	×	×	√
UPMC -Food101(Wang et al. 2015a)	90,840	101	×	√	×

Table 2.1: Summarization of food multimodal (or image-only) datasets.

restaurant geographic locations and 701 unique dish categories related. More general, cooking activity itself should also have a direct link to the categorization of dishes. (W. Susanto and Schiele 2012) records multi-view images of cooking activities in the kitchen. (Stein, S. and McKenna 2013) created 50 Salads using vision and accelerometers. They proposed three user adaptive models to robustly identify the difference between food/salad preparation activities. (Rohrbach and Amin 2012) propose a database of 65 cooking activities, continuously recorded in a realistic setting.

In this thesis, we propose a new dataset, UPMC Food-101, which contains about 101,000 images and textual descriptions for 101 food categories. Our dataset is different from the related works in following aspects: 1) our dataset is multimodal, which contains visual images and textual recipes. Note that as extra information, recipe is more informative than ingredients. For example, recipe describes the workflow, while ingredients not. 2) Different from the ETHZ Food-101, which is crawled from the professional food cooking websites, our dataset is crawled from the uncontrolled web search engine with a huge diversity among the instances.

2.2.2 *Food categorization*

Automatic food categorization is a key technology for various food-related research fields, such as cooking recipe retrieve (Salvador et al. 2017; Jingjing Chen 2016; Chen, Pang, and Ngo 2017; Matsunaga et al. 2015; Xie, Yu, and Li 2010), food recording (Aizawa and Ogawa 2015; Beijbom et al. 2015), food balance analysis (Aizawa et al. 2013; Kitamura et al. 2010; Christodoulidis, Anthimopoulos, and Mougiakakou 2015), food calorie estimation (Myers et al. 2016; Pouladzadeh, Yassine, and Shirmohammadi 2015; Pouladzadeh, Shirmohammadi, and Al-Maghrabi 2014; Miyazaki, Silva, and Aizawa 2011; Wu and Yang 2009), etc. Various methods are proposed by considering the specialty of food images or directly model this problem as generic image classification. Given an image of a dish, image-based food categorization maps a food image to a category of the dish. Various machine learning based image classification techniques are proposed to learn this mapping relation (Yang et al. 2010; He, Kong, and Tan 2016; Farinella, Moltisanti, and Battiato 2014; Hoashi, Joutou, and Yanai 2010; Matsuda and Yanai 2012). Especially, deep learning strategies shed light on this problem by learning more discriminative features (Jingjing Chen 2016; Keiji Yanai 2015; Kagaya, Aizawa, and Ogawa 2014; Kawano and Yanai 2014b). However, food categorization from an image remains a difficult problem because of the diversity of textures and a complicated mixture of elements (Oliveira et al. 2014). As a complement information of images, multimodal data such as ingredients (Jingjing Chen 2016; Min et al. 2016; Zhou and Lin 2016; Su et al. 2014; Yang et al. 2010), recipe text (Min et al. 2016; Jack Hessel 2015), restaurant geolocalization (Herranz, Jiang, and Xu 2017; Xu et al. 2015b; Bettadapura et al. 2015; Luis Herranz 2015) are exploited to build more robust classification systems. We shown in Table 2.2 the classification scores of parts of methods introduced above.

2.2.3 *Food-related multimedia applications*

Food-related multimedia applications are useful in many aspects. As an example, through a food image recognition system people can identify

	ETHZ Food-101	UEC FOOD 100	UEC FOOD 256	VIREO Food-172
(Kawano and Yanai 2014b)	- (-)	72.26 (92.00)	- (-)	- (-)
RF (Bossard and al. 2014)	50.76 (-)	- (-)	- (-)	- (-)
(Keiji Yanai 2015)	70.41 (-)	78.77 (95.15)	67.57 (88.97)	- (-)
(Myers et al. 2016)	79 (-)	- (-)	- (-)	- (-)
(Liu et al. 2016)	77.4 (93.7)	77.2 (94.8)	63.8 (87.2)	- (-)
(Hassannejad et al. 2016)	88.28 (96.88)	81.45 (97.27)	76.17 (92.58)	- (-)
(Jingjing Chen 2016)	- (-)	82.12 (97.29)	- (-)	82.06 (95.88)

Table 2.2: top-1 (top-5) classification accuracy (%) of competitive methods on large-scale food datasets.

the calorie, nutrient content, allergic ingredients, cooking methods, etc. For training such a system, sufficient number of labels of food images are required. An expert-based solution is performed by nutrition researchers or mechanical turk (Noronha et al. 2011). However, as data volume and user number increases, the human-based method faces bottlenecks of processing speed. (Bolaños, Garolera, and Radeva 2013) proposes an alternative method using active learning strategy for reducing the number of images for labelling by experts. It is also better to use the applications in a real-time environment. For pursuing this objective, (Kawano and Yanai 2015; Kawano and Yanai 2014c) proposes Foodcam. Foodcam is a automatic food image recognizer running on a smartphone. The local computation is achieved by reducing the amount of weight vector of the classifier. Also, (Oliveira et al. 2014) proposed a lightweight system to recognize prepared meals by segmentation and classification. More generic food-related multimedia applications are proposed when not considering these constrains. (Aizawa and Ogawa 2015) proposes daily food intake FoodLog. User creates a food log by uploading a photo taken by the phone. FoodLog can categorize the image into five categories and estimate its calorie by image content. Based on FoodLog, (Amano et al. 2014) shows that very small numbers of words are satisfactory to describe the majority of the record in the FoodLog system. Open Food System [1] aims at inventing new smart cooking appliances, with the ability to monitor cooking settings automatically for optimal results and preserve the nutritional value and

1 http://www.openfoodsystem.fr/

organometallic qualities of cooked foods. The Technology Assisted Dietary Assessment (TADA) project of Purdue University (Khanna and al. 2010) aims at developing a mobile food recorder, which can translate dietary information to an accurate account of daily food and nutrient intake. Food category classification is an indispensable ingredient in all these applications.

2.3 EYE-TRACKING RESEARCH

2.3.1 *Eye-tracking history*

Eye-tracking research has a long history and can date back to early 19^{th} century. It is a subject originated from the psychology research. At that time, researchers attempts to observe eye movements by direct observation. In 1879, Louis Émile Javal observed that reading does not involve a smooth sweeping of the eyes along the text, which is contrary to what people assumed for a long time. In 1908, Edmund Burke Huey (Huey 1908) defines the eye movement as a series of short pause *fixations* interrupted by rapid displacements *saccades*. From the two basic categories of gaze, people try to understand the meaning behind the gaze pattern. In the mid 20^{th} century, A. L. Yarbus (Yarbus 1967) showed that the task given to a subject has a very large influence on the subject's eye movement. Taking Fig. 2.3 for instance, empirical observations on these patterns demonstrate that the task largely influences the gaze pattern. A. L. Yarbus makes an assumptions that eye movement reflects the human thought processes. In 1980, Just and Carpenter (Just Marcel A. and Carpenter Patricia A. 1980) made further research for supporting the eye-mind relationship: there is no appreciable lag between what is fixated and what is processed. This hypothesis implies that when a human fixates on an area of an image, the information being processed derives from this area, rather than the areas been seen or to be seen. This hypothesis is challenged by (Michael Posner 1980), which claims that the cognitive precessing does not always relate to where the eye has been looking.

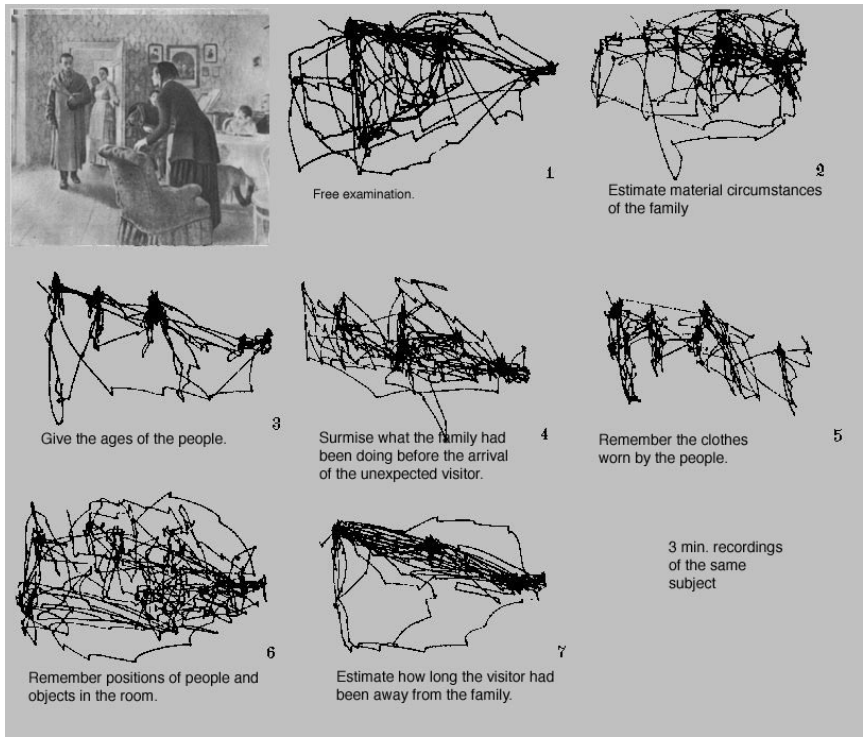


Figure 2.3: The classical task-influenced gaze pattern experiment conducted by A. L. Yarbus (Yarbus 1967). In this experiment, subjects are given specific tasks before observing the famous painting Unexpected Visitors (1888, by Ilya Repin). Empirical observations on these patterns demonstrate that the task largely influences the gaze pattern.

With the advent of small, highly accurate and low-cost systems, eye tracking is now being rapidly adopted in devices and applications, both to enhance computer interaction and to understand human behavior. This technique has been used in areas like augmentative alternative communication, gaming, health care, market research, performance assessment (athletics, online course), neuroscience, driver assistance systems, assessment of user experience, etc.. The psychological issues, like predicting the gazes, salience or scan-paths (Mathe and Sminchisescu 2013; Vig, Dorr, and Cox 2012; Sattar et al. 2015), eye-mind relation (Klami et al. 2008; Yun et al. 2013), are also inspiring the evolution of computer vision systems.

2.3.2 *Eye-tracker devices*

Eye-tracking research requires a precise eye-tracker for recording the eye movements. In the early 1900s, an eye-tracker was built using a contact-lens-like device. The lens was connected to an aluminum pointer that moved in response to the movement of the eye. Since the lens contact directly the human eyes, this kind of device is call *intrusive eye-tracker*. It is obvious that the intrusive eye-tracker will affect the recorded data. The first *non-intrusive* eye trackers was invented in the 1950s. They reflected beams of light onto the eye and then recorded them on film. Another method was to use simple 8-mm film to track eye movement by filming the subject through a glass plate, on which the visual problem was displayed. Since then, eye-tracker product is becoming more and more portable and precise. Nowadays, there are a number of eye-tracking hardware companies, including Tobii, SensoMotoric Instruments (SMI), EyeLink, etc. During the research of this thesis, we use the model X2-30 of Tobii, which can be hidden under the screen of a laptop. As shown in Fig. 2.4, this eye-tracker can be placed just under the 11.6-inch screen of a laptop, which does not disturbs the subject during the experiments.

There are two steps before using the X2-30 to collect eye movement informations: calibration and tracking. Calibration is a step for measuring characteristics of the user's eyes and uses them together with an internal, physiological 3D eye model to calculate the gaze data. This model includes

Figure 2.4: Tobii X2-30 eye-tracker appearance.

information about shapes, light refraction and reflection properties of the different parts of the eyes (e.g. cornea, placement of the fovea, etc.). During the calibration the user is asked to look at specific points on the screen, also known as calibration dots. During this period several images of the eyes are collected and analyzed. The resulting information is then integrated in the eye model and the gaze point for each image sample is calculated. Tracking step is shown in Fig. 2.5: Human eyes are illuminated by a light source then reflect the light, a camera then captures an image of the eye showing these reflections. By identifying the reflection on the cornea and in the pupil, and combining with other geometrical features of reflections, eye-tracker is able to calculate the gaze direction. The intersection point of gaze direction and object plane is the gaze point. As our objective is to exploit eye-tracking features for image recognition, we limit our concerns on the output data of eye tracker. They are:

1. Time stamp: This is a value that indicates the time when the information used to produce the gaze data packet was sampled by the eye tracker.

2. 3D eye position: The eye position is provided for the left and right eye individually and describes the position of the eyeball in 3D space.

3. 3D relative eye position: The relative eye position is provided for the left and right eye individually and gives the relative position of the eyeball in the track box volume as three normalized coordinates.

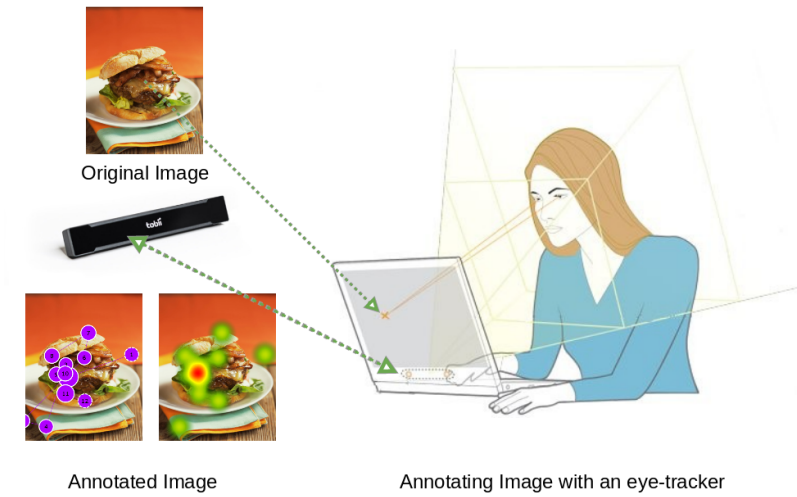


Figure 2.5: The working principle of eye-tracker.

4. 3D coordinates of gaze point: The 3D gaze point (or gaze position) is provided for the left and right eye individually and describes the position of the intersection between the calibration plane and the line originating from the eye position point with the same direction as the gaze vector.

5. 2D coordinates of gaze points: The 2D gaze point is provided for the left and right eye individually. It is conceptually the same as the 3D gaze point, but expressed as a two-dimensional point on the calibration plane instead of as a point in 3D space.

6. Pupil diameter: The pupil diameter data is provided for the left and the right eye individually and is an estimate of the pupil size in millimeters.

7. Validity of records: An estimate of how certain the eye tracker is that the data given for an eye really originates from that eye.

The calibration plane in our work is the plane of computer screen. 3D position informations of eyes are important for locating eyes and describing

the spatial positional relationship between eyes and calibration plane. It's the data for tracking eyes rather than eye-tracking data. Gaze information describes what we really looked. According to the definition, 3D gaze position can be converted to 2D gaze position. Based on this analysis, we think 2D gaze position is sufficient for describing where we are currently looking. Moreover, as gaze positions are represented in a series of time, the path of gaze points can be painted without no saccade points, which means quick, simultaneous movement of both eyes between two phases of fixations. Fixation and saccade are two kinds of eye movement mostly researched in the history of eye-tracking research. In Fig. 2.6, we demonstrate the two types of gaze, with circle represents the fixation and the line represents the saccade.

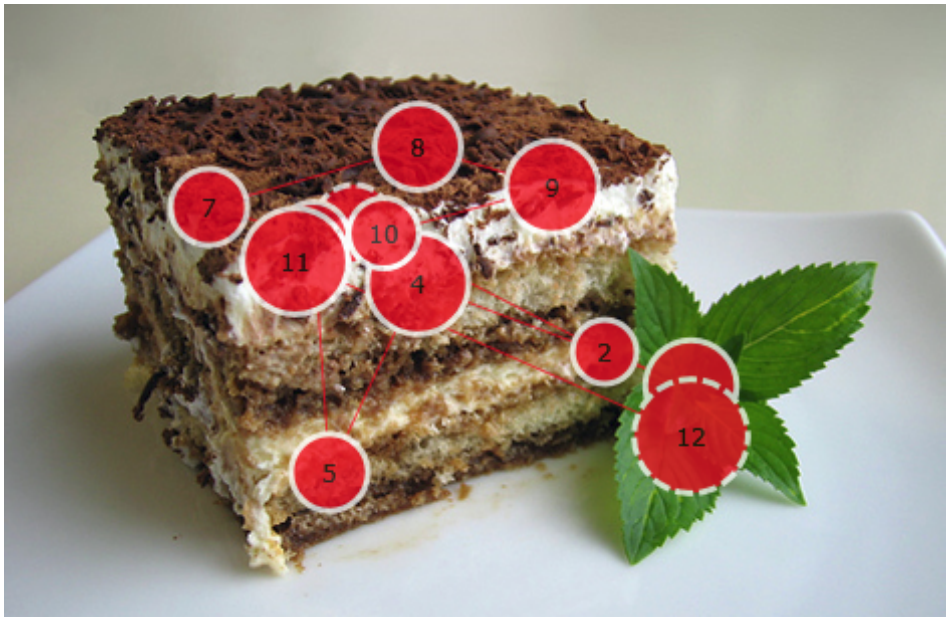


Figure 2.6: Fixations (annotated as round points) and saccades (annotated as the lines). The radius of fixation represents the duration of the fixation. The number indicates the order of fixations.

2.3.3 *Eye-tracking in computer vision*

In this thesis, we focus on building automatic visual understanding system. In computer science field, from 1980s to now, eye-tracking technology was mainly applied in the field of human-computer interaction. The applications include eye-controlled role-play video game, wink-controlled photo taking program, eye-movement controlled computer designed for particular user (Jacob 1995), web interface attention analysis for designing layout (advertisements, title, etc.) (Jacob and Karn 2003). We refer our reader to a recent review of human-machine interaction using eye-tracking (Majaranta and Bulling 2014) of this topic.

In fact, human-computer interaction and our objective overlap each other within *attention*. In the human-computer interaction, human *pays* the attention to interact with the machines, while in visual understanding system machine can *borrow* the human attention or *generate* its own attention to assist the learning process. The *borrow* scenario relies on the top-down attention while the *generate* scenario is often referred as bottom-up attention (Le Meur et al. 2006; Le Meur and Le Callet 2009). In this thesis, we mainly study the top-down attention, which is acquired directly from human. This kind of attention is highly related to main objects in image (Wang, Chandler, and Le Callet 2010; Ramanathan, Yanulevskaya, and Sebe 2011; Yun et al. 2013). For capturing the human attention, we intend to use the eye-tracker (subsection 2.3.2). However, as we know, human gaze does not equal the attention as there exists overt attention and covert attention (Le Callet and Niebur 2013), where the latter indicates the attention is not always coherent with the center of gaze (Le Callet and Niebur 2013), *e.g.* a car driver who fixates the road while simultaneously and covertly monitoring road signs and lights that appear in the retinal periphery. In this thesis, we roughly treat attention as gaze because our proposed models are able to learn from such kind of weak supervision.

In computer vision research field, people usually use gaze for solving many problems, including image/video quality assessment (Ninassi et al. 2007; Zhang et al. 2016; Ninassi et al. 2009) inferring subject's searching task (Sattar et al. 2015; Borji and Itti 2014; Haji-Abolhassani and Clark 2014;

Zelinsky, Peng, and Samaras 2013), action recognition (Bulling et al. 2011; Steil and Bulling 2015; Ge et al. 2015; Mathe and Sminchisescu 2015), gaze path prediction (Hacisalihzade, Stark, and Allen 1992; Kashlak et al. 2017), saliency prediction (Pan et al. 2016; Kruthiventi et al. 2016; Mathe and Sminchisescu 2013; Wang et al. 2013a), segmentation (Shcherbatyi, Bulling, and Fritz 2015; Walber, Scherp, and Staab 2013; Karthikeyan et al. 2013; Papadopoulos et al. 2014; Ramanathan et al. 2010; Mishra, Aloimonos, and Cheong 2009), object detection (Fathi, Li, and Rehg 2012; Yun et al. 2013). In video analysis, since subjects tend to watch at the moving objects, gaze are also widely used to localize important objects (Karthikeyan et al. 2015; Shapovalova et al. 2013; Damen, Leelasawassuk, and Mayol-Cuevas 2016; Xu et al. 2015a). A group of researches are related to visual preference-based image retrieve. (Papadopoulos, Apostolakis, and Daras 2014) formalizes the visual preference in a binary classification problem based on gaze features. Pinview (Hussain et al. 2014) is a image retrieve system based on the user's visual preference represented by either mouse click or gaze. Recently (Sattar, Bulling, and Fritz 2016) proposes a gaze pooling layer, which integrates gaze information into CNN-based architectures as an attention mechanism. Interestingly, people has trained a eye-tracker from the webcam using deep learning strategy (Krafka et al. 2016).

Also, gaze features are appealing since they can be generated by humans at almost zero-cost when performing a recognition task. Collecting gazes is more user-friendly and less time-consuming than collecting traditional annotations. According to the published works, it takes about 1 second to collect gazes for one image (Papadopoulos et al. 2014), comparing to 26s for drawing a bounding-box (Su, Deng, and Fei-Fei 2012) and 15-60 min for labeling the segmentation mask for an image (Pushmeet Kohli and L'ubor Ladický and Philip H.S. Torr 2009). Gaze is often converted to fixation density maps for smoothing the gaze distribution (Engelke et al. 2013). For different purposes, people design different collection protocols to acquire gazes (Lopez et al. 2015; Papadopoulos et al. 2014; Mathe and Sminchisescu 2013; Karthikeyan et al. 2013). The collection protocols can be grouped into two main groups: task-driven and free-viewing. Task-driven

means the annotators are given a specific semantic to look at, *e.g.* dog or a group of actions. Free-viewing means the annotators view the image freely. In this chapter, we use two task-driven datasets (Papadopoulos et al. 2014; Mathe and Sminchisescu 2013).

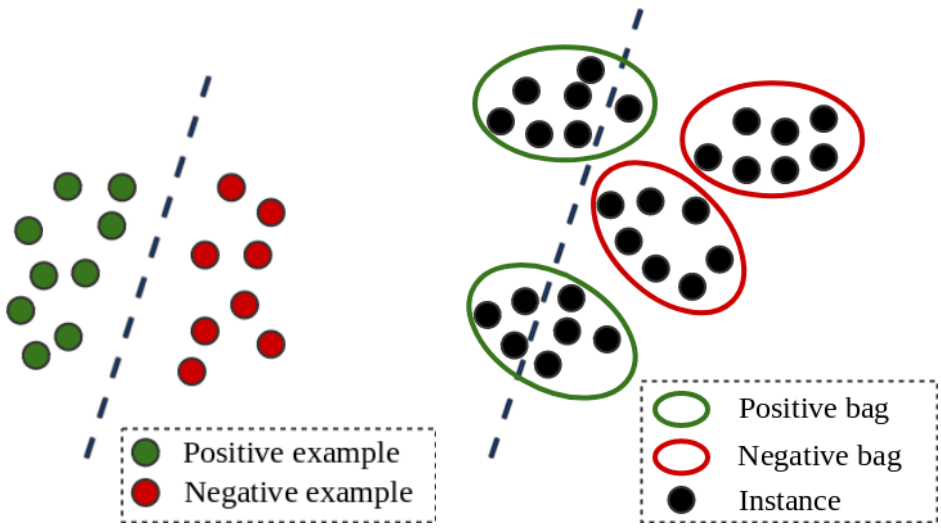
To acquire gaze annotations for different applications, people design various collection protocols (Lopez et al. 2015; Papadopoulos et al. 2014; Mathe and Sminchisescu 2013; Karthikeyan et al. 2013). The collection protocols can be grouped into two categories: *task-driven* and *free-viewing*. Task-driven means the annotators are given a specific semantic to look at, *e.g.* a dog. Free-viewing means the annotators view the image freely without specific purpose. As an example of *free-viewing*, Lopez et al. (Lopez et al. 2015) expose simultaneously two images on the screen for evaluating the annotator's visual preference. The aim of this protocol is to collect the gaze features of left and right image for classifying the visual preference. Papadopoulos et al. (Papadopoulos et al. 2014) use an instantiation task-driven protocol. Specifically, this protocol first group image categories into visual-similar pairs. Then the annotation interface exposes to the annotator one image from a selected pair. The annotator should make a decision on the category of the image. The advantage of this protocol is that it does not need the target-absent image to avoid guess, which further reduces unnecessary labeling time. Similarly, Mathe et al. (Mathe and Sminchisescu 2013) annotate two concepts: *actions* and *context*. One image is exposed to the annotator. Then the annotator is told to find all the actions in the image. Since then, gaze in one image are related to all categories. Gilani et al. (Gilani et al. 2015) use a similar protocol as (Mathe and Sminchisescu 2013). But additionally, they have an extra free-viewing protocol for comparing the internal connection with the task-driven protocol. In this thesis, we propose a new dataset, UPMC-G20, with gaze annotation using a similar task-driven protocol as in (Papadopoulos et al. 2014). This dataset is based on the large-scale food-related dataset UPMC Food-101 (Wang et al. 2015a). We make this dataset because food is also a complicate object made up of various ingredients and often with several common backgrounds. The detail of UPMC-G20 is described in section 5.3.

2.4 WEAKLY SUPERVISED LEARNING

Classifying clutter image can be difficult because the useful object information is hidden under a lot of noise. In order to improve the accuracy of recognition, an intuitive idea is to select the image area associated with the target semantics to classify. However, collecting full annotations for all the images in a large dataset is an expensive task: whereas several millions of images annotated with a global label are nowadays available, only limited accurate bounding box annotations exist (Matthew Blaschko, Pawan Kumar, Ben Taskar 2013). This observation makes the development of Weakly Supervised Learning (WSL) models appealing. WSL is an attractive learning strategy because it can mine more local information with less intensive annotations with respect to the supervised learning.

2.4.1 *Multiple Instance Learning (MIL)*

Multiple Instance Learning (MIL) is one of the main paradigms for training WSL models. The term of Multiple Instance Learning was first proposed by (Dietterich, Lathrop, and Lozano-Pérez 1997) for predicting drug activity. In this problem, biochemists can produce a set of molecules and identify whether the set is qualified to make a drug or not. But they do not know which conformations of the molecules are responsible for the drug activity. This setting is different from the supervised learning where all the examples are labeled (as shown in Fig. 2.7a). MIL is formalized as shown in Fig. 2.7b: An example is first represented as a labeled *bag*, which contains a number of instances without labels. MIL supposes that there is at least one positive instance in the positive bag, while all instances in the negative bag are negative. This hypothesis links the latent instance label and the ground-truth bag label. Based upon this formulation, a lot of MIL algorithms are proposed, for example, Diverse Density (Maron and Lozano-Perez 1998a; Maron and Lozano-Perez 1998b), Citation-kNN and Bayesian-kNN (Wang and Zucker 2000), EM-DD (Zhang and Goldman 2001), MI kernels (Gärtner et al. 2002), multi-instance ensembles (Zhou and Zhang 2003) and neural network-based methods (Wang et al. 2016; Zhang and Zhou 2004) Recently,



(a) Supervised learning (b) Multiple-instance learning

Figure 2.7: Supervised learning vs MIL : in supervised learning all the examples are labeled whereas in MIL only the bags are labeled, i.e. the instance labels are unknown. The blue dotted line shows the separator learned by the classifier.

WSL is widely applied into many computer vision fields, including object detection(Ren et al. 2016; Felzenszwalb et al. 2010; Wang et al. 2015b; Shen et al. 2016), scene recognition (Juneja et al. 2013; Sun and Ponce 2013; Pandey and Lazebnik 2011) and dictionary learning (Wang et al. 2013b; Shrivastava et al. 2015). The Deformable Part Model (DPM) (Felzenszwalb et al. 2010) has been extremely popular due to its excellent performances for weakly supervised object detection. The DPM learns object part filter and applies on sub-region of an image to get a response and consider the spatial prior. In the DPM, the Latent Support Vector Machine (LSVM), which is also known as MI-SVM (Andrews, Tsochantaridis, and Hofmann 2002), describes the MIL assumption in the form of classical SVMs. One challenge with LSVM is due to the introduction of latent variables, which makes the resulting optimization problem non-convex. When using sliding window

approaches for generating the candidate regions, the size of the latent space becomes enormous. To overcome this issue, incremental exploration strategies have been proposed in (Durand et al. 2014; Russakovsky et al. 2012b; Bilen, Namboodiri, and Gool 2014; Kumar, Packer, and Koller 2010). Advanced instance selection methods model the relations among the instances by graph model (Zhou, Sun, and Li 2009; Deselaers and Ferrari 2010) and recursive neural network (Garcez and Zaverucha 2012). Finally, recent works focus on enriching the prediction function, by using several (top) instance scores instead of using a single max (Li and Vasconcelos 2015), or by incorporating negative evidence (Azizpour et al. 2015; Durand, Thome, and Cord 2015; Durand, Thome, and Cord 2016; Durand et al. 2017). Under the weakly supervised circumstance, MIL is widely used in object localization (Ren et al. 2016) (MI-SVM like with positive bag split) There are several comprehensive reviews about MIL (Zhou 2004; Babenko 2009; Foulds and Frank 2010; Amores 2013; Carbonneau et al. 2016). Recently, a book talking about MIL is published (Herrera et al. 2016).

2.4.2 *WSL eye-tracking research*

Recently, attempts have been devoted to incorporating gazes as weak supervision signals (Fathi, Li, and Rehg 2012; Papadopoulos et al. 2014; Ge et al. 2015) for improving the performance of classification or segmentation systems. Relevant to our proposed models, Mathe et al. (Mathe and Sminchisescu 2014; Mathe, Pirinen, and Sminchisescu 2016) proposes using reinforcement learning to find a latent space sampling policy from gaze. This method is efficient at the cost of prediction accuracy. Karthikeyan et al. (Karthikeyan et al. 2013) proposes to train a face and text detector from only gaze information. Although this work does not use image features, it still requires bounding boxes to segment out face and text regions. Shcherbatyi et al. (Shcherbatyi, Bulling, and Fritz 2015) integrates gaze into Deformable Part Model for selecting one relevant object location. Their model require gaze annotations for test. Shapovalova et al. (Shapovalova et al. 2013) focuses on WSL recognition by penalizing region selection with

gaze. However, the gaze information is not sufficiently exploited because only positive examples are penalized with gaze.

2.5 CONCLUSION

In this chapter, we have reviewed three research directions related to this thesis, namely image recognition technology, eye-tracking technique and weakly supervised learning. In section 2.1, we first introduced two of the most commonly used image representations: hand-crafted and ConvNet-based. A key success of the ConvNet is that the learned representations on ImageNet are both discriminative and generic, so they can be efficiently transfered to other datasets 2.1.2. As we concern more about the food dataset recognition in this thesis, in the section 2.2, we study current food datasets, food image recognition methods and applications. In this specific data scenario, we also observe that ConvNet-based representation is the state-of-the-art. In Chapter 3, we present a large scale food-related multi-modal dataset. This dataset was the first that contains a food image/recipe pair at the time of publication. In this work we combine the state-of-the-art computer vision and natural language techniques together for building a more robust recognition system.

The objective of our research is to reduce the labeling cost when training an image classifier. For achieving this goal, we carry out our research from two aspects: low-cost gaze annotation and weakly supervised learning. On one hand, the hypothesis of the eye-mind relationship connects the human eye perception signals and the understanding of the image content, which is still an important theoretical basis for modern eye-tracking based machine vision method 2.3.1. From the observation-based invasive ancient devices to the modern precise non-invasive eye-tracker, the human eye tracking results become increasingly trustworthy 2.3.2. In our research, we use the eye-tracking data in a pragmatic manner. On the other hand, weak supervised learning (WSL) has gained widespread attention due to the efficient use of data labels. The most popular approach for WSL in computer vision is Multiple-Instance Learning (MIL) (Subsection 2.4.1). MIL is a binary classification problem where a class label is assigned only

to a bag of instances, indicating the presence/absence of positive instances. The standard MIL assumption is: a bag is positive if it contains at least one positive instance, and negative if it contains only negative instances. This assumption infers the latent instance label from the ground-truth bag label, so it does not requires rich instance-based label. For benefiting the goodness from both gaze annotation and MIL, in Chapter 4 and 5, we introduce two gaze-based MIL models for simultaneously classifying and localizing discriminative parts of objects. For consolidating the robustness of our model, we annotate part of the our food dataset with an eye-tracker and observe a consistent performance gain across various benchmark datasets.

3

MULTIMODAL FOOD RECOGNITION AND APPLICATION

ABSTRACT

In this chapter, we introduce a large scale multimodal food-related dataset: UPMC Food-101. The objective of building this dataset is to experiment the state-of-the-art image representations and models into a real application exploring web resources: finding good pictures for recipes. For building the UPMC Food-101, we take into account of the crawling, cleaning and ranking data procedures. Based on this dataset, we perform deep analysis of category classification and recipe retrieve using the visual and/or textual information. We also present experiments with text-based embedding technology to represent the relations among food words in a semantical continuous space. We compare our dataset with another food dataset: ETHZ Food-101 (Bossard and al. 2014). We revisit the data collection protocols of ETHZ Food-101 and carry out domain adaptation experiments to highlight the similarities and differences between both datasets. We then build a web-based application based on the UPMC Food-101 using deep learning models, which allows querying an image and retrieving the most relevant recipes from our dataset.

The work in this chapter is published as:

- Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso (2015a). "Recipe recognition with large multimodal food dataset." In: *IEEE International Conference on Multimedia & Expo Workshops*, pp. 1–6.

3.1 INTRODUCTION

In this chapter, we focus on building an automatic system for recipe recognition. For validating our systems, we first propose a new large-scale multimodal dataset: UPMC Food-101. UPMC Food-101 contains about 100,000 recipes out of 101 food categories. This dataset is collected from the web by satisfying some constraints, which we describe in section 3.2. Each item in this dataset is represented by one image and its original HTML page content. Based on this dataset, we conduct extensive experiments to explore the properties in terms of classification performance and recipe semantic relation. Specifically, we first compare both shallow and deep features of images on this dataset. Related to our work, (Yang et al. 2010) extracts pixel pairs features for describing the ingredient spatial distribution on a food image. (Bossard and al. 2014) uses a random-forest for mining discriminative parts clusters in the food images for training an SVM. (Aizawa et al. 2013; Kitamura et al. 2010; He, Kong, and Tan 2016) uses SVMs for generic food categorization. Considering deep models, (Kawano and Yanai 2014b) performs a late fusion of deep convolutional features and conventional hand-crafted image features, which outperforms both features. (Keiji Yanai 2015) fine-tunes Alexnet, (Myers et al. 2016; Liu et al. 2016) fine-tunes the GoogleLeNet (Szegedy et al. 2015) while (Hassannejad et al. 2016) fine-tunes the Google Inception-v3 (Szegedy et al. 2016). (Jingjing Chen 2016) adapts a multi-task VGG-16 (Simonyan and Zisserman 2015) for classifying both food categories and ingredient categories, which benefits both tasks. Furthermore, count-based textual features, i.e. TF-IDF, and embedded-based textual features, *e.g.* word vector (Mikolov and all 2013), are coupled with image features for highlighting the complementarity of multimodal data.

This chapter is organized as follows. In Sec. 3.2, we detail the motivation, method, and originality of our dataset. In Sec. 3.3, we perform classification experiments at a large scale to evaluate visual and textual features along with their fusion. In Sec. 3.4, we propose further statistics to highlight dataset characteristics and comparison with another popular large scale food dataset ETHZ Food-101 (Bossard and al. 2014). In Sec. 3.5,

we demonstrate the interest of these recognition technologies coupled with UPMC Food-101 in a mobile search application. We conclude this chapter in Sec. 3.6.

3.2 UPMC FOOD-101 DATASET

3.2.1 *Data Collection Protocol*

To create a real world and challenging dataset with multimodal data, we use the Google Image search. Unlike controlled sources, using the Web search engine allows to explore recipes that are potentially deeply buried in the world wide web. Similarly, (Kawano and Yanai 2014a) explores the Web resources to extend their initial UEC Food-100 dataset. It is also interesting to note that the past approaches (Schroff, Criminisi, and Zisserman 2011) using Google search engine to obtain images for classification tasks have reported around 30 percent of precision level on some of collected images (in 2006). We observe that the results returned by Google Image search in 2014 for textual queries related to food images are more relevant with low level of noise. This is explained by the large improvement in the field of searching and page ranking algorithms since 2006. Based on these preliminary findings, we decide to create our database by querying Google image search with 101 labels taken from the ETHZ Food-101 dataset (Bossard and al. 2014) along with an added word ”recipes”. We added the word ”recipes” to each label before passing the query to Google for two reasons:

- As we are interested in recipe recognition, adding ”recipes” word after the labels, for example, ”hamburger recipe”, returns more focused information about ”how to make hamburgers” rather than other topics like ”where to eat hamburgers” or ”Hamburger is junk food” in the textual form.

- We observed that adding ”recipes” to our queries helps decreasing the noise level a little further in the returned images. For example, a simple ”hamburger” in search engine could return some thing like

"hamburger menu icon" or "hamburger-like evening dress", which are far from our expectations.

3.2.2 *Crawling Google: engineering details*

Retrieving images from the Google search engine by keyword is straightforward: crawling Google results using a script. It should be pointed out here that as Google results use AJAX, directly crawling and extracting the links for images and seed pages from the HTML page source will return only few results. Also, as Google only returns 20 results per AJAX call, we have to iteratively submit our query by changing the starting index. Through this script links for up-to 1000 images/html pages per query can be collected as the absolute maximum number of results returned by Google for each query is 1000.

Once we collected the links of images and their respective HTML seed pages, we just directly downloaded the images and HTML data from the specific URLs. One point to take notice of here is that sometimes there is a failure in data collection from some of these URLs due to the following reasons: 1) The Data has been moved. or 2) Script/text based response is banned by that particular website.

3.2.3 *Content of UPMC Food-101*

We then collect the first 1,000 images returned for each query and remove any image with a size smaller than 120 *pixels*. In total, UPMC Food-101 contains 101 food categories and 90,840 images, with a size range between 790 and 956 images for different classes. Fig. 3.1 shows representative instances of all 100 categories. Due to no human intervention in grasping these data, we estimate that each category may contain about 5% irrelevant images for each category. 3 examples of "hamburger" class are shown in Fig. 3.2. We notice that adding the keyword "recipes" results in taking into account ingredient or intermediate food images. Determining whether these images should be considered as noise or not, directly depends on the specific application. Additionally, we save 93,533 raw HTML source pages

along with the embed images. The reason that we don't have 101,000 HTML pages is that some pages are not available. The number of the images that have text is 86,574.

Figure 3.1: 100 samples out of 101 categories of UPMC Food-101 dataset.

3.2.4 *Comparison with ETHZ Food-101*

The food dataset ETHZ Food-101 (Bossard and al. 2014) has been recently introduced. 101,000 images for 101 food categories have been collected from a specific website (*e.g.* www.foodspotting.com). The labels of food

(a) Correct (b) Ingredient (c) Noise

Figure 3.2: Example images within class "hamburger" of UPMC Food-101. Note that we have images completely irrelevant with hamburger like Figure 3.2c, as well as hamburger ingredient like Figure 3.2b, which reflects the real distribution of the resutls returned by the search engine.

Dataset	class num	image num per class	source	Data type
UPMC	101	790 - 956	various	text&image
ETHZ	101	1000	specific	image

Table 3.1: UPMC Food-101 and ETHZ Food-101 dataset content.

categories were chosen from the top 101 most popular dishes on the mentioned website.

We have used the same class labels as ETHZ Food-101 for our dataset. In Table 3.1, general statistics on both sets are reported. The main difference comes from the data collection protocols. Since our data is collected directly from a search engine with automatic annotations, whereas ETHZ Food-101 dataset images were collected from a specific website, which contains manual annotated images uploaded by humans, leading to less number of false positive/noise in ETHZ Food-101 than in UPMC Food-101. As the three examples of "hamburger" class show in Fig. 3.3, ETHZ Food-101 ensures images irrelevant with food categories are mostly excluded from this dataset. Moreover, there was no textual data provided with images in ETHZ Food-101. However, to classify between two variants of the same food categories, text can help a lot. We explore visual and text classification in the next section.

Figure 3.3: Example images within class "hamburger" of ETHZ Food-101. All these images have strong selfie style as they are uploaded by consumers. Although some background noise (human faces, hands) are introduced in images, it ensures images out of food categories are excluded from this dataset.

3.3 CLASSIFICATION RESULTS OF UPMC FOOD-101

In the following subsections we run several classification algorithms by using visual information, textual information and the fusion, to make quantitative descriptions of our dataset. The results are shown in Table 3.2. A unified training and test protocol is applied for both visual and textual tests, in order to evaluate and compare the performances with minimal extra factors. The protocol is as follows: we split out the examples, which have both image and text, then randomly select 600 training examples for each category to train a one-vs-rest linear SVM (Fan and al. 2008) with $C = 100$, the remaining examples are for test. We evaluate our results by averaging accuracy over 10 tests, where accuracy is defined as $\frac{\#(true\ positives)}{\#(test\ examples)}$.

Visual				Textual	Fusion
BoW	Bossanova	OverFeat	VGG-16	TF-IDF	TF-IDF + VGG-19
23.96	28.59	33.91	40.24	82.06	85.10

Table 3.2: Top-1 Classification results (Ave. accuracy %) on UPMC Food-101 for Visual, Textual and fusion features.

3.3.1 *Visual Feature Classification*

3.3.1.1 *Bag-of-Words Histogram (BoW) + SIFT*

We represent images as Bag-of-Words histogram with a spatial pyramid as our first baseline. In detail, we first proportionally resize images, which has a size larger than 300 *pixels*, then extract mono-scale SIFT with window size 4 and step size 8, 1024 word visual dictionary, soft coding and max pooling with 3 level spatial information. This baseline obtains an average accuracy 23.96%.

3.3.1.2 *Bossanova Image Pooling Representation*

Bossanova (Avila et al. 2012) reinforces the pooling stage of BoW by considering distance between a word and a given center of a cluster. As Bossanova only modifies the pooling stage, we can reuse the same coding setting as BoW. In our experiment, 2 bins are used in the quantization step to encode the distances from sifts to clusters, BoW is concatenated with vector histogram with no scaling factor, we set range of distances per cluster to $[0.4, 2.0]$, for each word we consider 10 neighbors. This method results in an average accuracy of 28.59%, which constitutes an improvement of 19.37% over the BoW model.

3.3.1.3 *Deep Feature Models*

CNN deep feature is the state of the art in many image recognition challenges. Deep feature is more expressive than hand-crafted image features In our experiment, we first adopt the "fast network" pre-trained model of OverFeat[3] as the feature extractor. of a given image. We get an average accuracy of 33.91%.

This result is interesting because the OverFeat CNN was trained on 1,000 class dataset ILSVRC2012, which contains very few images of food categories (French fries, few images of waffles etc). Even after having been trained on few food images, the OverFeat CNN produces very good deep

3 http://cilvr.nyu.edu/doku.php?id=software:overfeat:start

features, which outperform the standard Bossanova baseline in the context of classification.

(Simonyan and Zisserman 2015) pushes CNN network to $16 - 19$ weight layers, which is about twice deeper than the previous work. In our experiment, we use the pre-trained model "vgg-16"[4] to extract features. This model is also trained on ILSVRC2012, so it is comparable with the OverFeat. The `4096d` output after the ReLU activation of the `FC7` layer is used as the feature description. We finally achieve an accuracy of 40.21% over our dataset with these features.

We further retrain from scratch and fine-tune deep models on the UPMC Food-101 (Cadène, Thome, and Cord 2016). We adopt the same architecture, namely the OverFeat and the vgg-16 as orginal models. For fine-tuning the models, we replace the output number 1000 by 101 and retrain the last fully connected layers. The results are reported in Table 3.3.

OverFeat-r	OverFeat-f	VGG-16-r	VGG-16-f	InceptionV3-f
47.46 (69.37)	57.98 (78.86)	53.62 (74.67)	65.71 (82.54)	66.83 (84.53)

Table 3.3: Fine-tuning and learning from scratch classification results (top-1 (top-5) Ave. accuracy %) on UPMC Food-101. *-r: retraining model from scratch, *-f: fine-tuning model.

Comparing these results, we find that fine-tuning the pretrained models adapts best to the UPMC Food-101. We also observe that training a deep network from scratch largely outperforms directly extracting features from pretrained model, this is opposite to what we observe on the small dataset such as PASCAL VOC, which suffers from severe overfitting when training on a large network.

3.3.2 *Visual Domain Adaptation*

As another set of visual experiments, we perform knowledge transferring experiments over both datasets (ETHZ Food-101 and UPMC Food-101), namely learning the classifier model on one dataset and testing it on the

4 http://www.vlfeat.org/matconvnet/pretrained/

other one. This experiment aims at showing the different performances of UPMC Food-101 and ETHZ Food-101 when performing visual classification. In this experiment, we use very deep features. The results of the transfer learning experiments are shown in Table 3.4. The first two rows show the results of classification when training with the same number of examples (*e.g.* 600 examples for each class) of one dataset and testing on the rest of this dataset or on the whole of the other dataset, while the last two rows show the results of classification when training with all examples on one dataset and testing on the other dataset.

There are some interesting points that can be inferred from the results. The first one is that even though both datasets contain images for same food categories, they are very different from each other. This can be derived from the fact that there is a considerable difference of around 50% average accuracy when training on one dataset and testing on both datasets (first 2 rows in Table 3.4).

Second point that can be observed from the Table 3.4 is that training on part of UPMC Food-101 outperforms training on the whole UPMC Food-101 when testing on ETHZ Food-101 by a margin of 1.57%, while on the contrary, only a negligible difference (0.36%) for training on ETHZ Food-101 and testing on UPMC Food-101 is observed. This perhaps can be an indication of comparative noise levels in both datasets, UPMC Food-101 being more noisy.

train / test	UPMC	ETHZ
UPMC (600 examples)	40.56	25.63
ETHZ (600 examples)	25.28	42.54
UPMC (all examples)	-	24.06
ETHZ (all examples)	24.92	-

Table 3.4: Average accuracy of transfer models between UPMC Food-101 and ETHZ Food-101.

Note that our ETHZ deep results are not comparable with the CNN results in (Bossard and al. 2014) because they train deep features as we use a pre-trained CNN on ImageNet.

3.3.3 *Textual Feature Classification*

Since our raw textual data is in html format, we need some preprocessing in order to remove numerous noisy elements such as html tags, code, punctuations. Our foremost preprocessing is parsing content out from HTML pages by Python package html2text[5].

3.3.3.1 *TF-IDF*

TF-IDF (Term Frequency–Inverse Document Frequency) value measures the importance of a word w in a document D with respect to the whole corpus, where TF evaluates the importance of word in a document, and IDF evaluates the importance of a word in the corpus.

To represent a document with TF-IDF, we generate the dictionary by preprocessing words as follows: 1/ Stemming all words. For example, words like "dogs" and "sleeping" are respectively stemmed to "dog" and "work", 2/ Removing words with high frequency of occurrence (stop words) such as "the","is","in", 3/ Removing words occurred less than in 11 docs, 4/ Keeping stems with length between 6 and 18. After the pre-processing, 46972 words are left. We then form a dictionary $Dict_t$ using these words.

We calculate TF-IDF value for every word in document by formula $tfidf_{w,D} = tf_{w,D} \times idf_w$, with $tf_{w,D} = \frac{n_{w,D}}{\sum_k n_{k,D}}$, where $n_{i,j}$ is the frequency of word i appearing in document j, and $idf_w = \log \frac{|N|}{|\{j:w \in D_j\}|}$, where N is the total number of documents in the corpus, and $|\{j : w \in D_j\}|$ is the number of documents where the term w appears. TF-IDF value favors the words less occurred in corpus and more occurred in a given document D, and suppress the word in reverse case. A document can be represented by the TF-IDF value of all its words belonging to the dictionary $Dict_t$. We obtain 82.06% classification average accuracy on our dataset. Such a high score is partly due to the bias introduced by our data crawling protocol.

5 https://pypi.python.org/pypi/html2text

3.3.4 *Late Fusion of Image+Text*

We merge very deep features and TF-IDF classification scores by late fusion. The fusion score s_f is a linear combination of the scores provided by both image and text classification systems, as $s_f = \alpha s_i + (1 - \alpha)s_t$, where α is the fusion parameter in the range $[0, 1]$, s_i is the score from the image classifier and s_t is the score from the text. We select α by cross-validation over different splits of data and the final classification score is 85.1%, which improves 3.6% with respect to textual information alone and 109.8% with respect to visual information alone. Note that the classification scores were not calibrated prior to late fusion so that α does not depend on the relative accuracy of each source of scores.

3.4 QUALITATIVE ANALYSIS OF UPMC FOOD-101

In this section, we report further analysis of UPMC Food-101. We investigate the word vector representations (Mikolov and all 2013) for its strong semantic expressiveness. Transfer learning between UPMC Food-101 and ETHZ Food-101 is also analyzed.

3.4.1 *Word Vector Representation*

We first introduce how to extract word vectors, then explore some interesting features of this representation.

After parsing out the content of web pages, we concatenate all of them together to build a corpus for training a dictionary $Dict_v$ with word2vec (Mikolov and all 2013), which is a tool to efficiently compute vector representations of words. Words with an occurrence frequency less than 5 in the corpus are removed from $Dict_v$. This condition results in 137092 words, in which each word is described by a 200 dimensional feature vector. $Dict_v$ contains stop words and other noisy words, so we intersect $Dict_t$ and $Dict_v$, which creates a new dictionary $Dict$ containing 46773 words.

On the other hand, each document is first preprocessed by the tool html2text, then represented by the element-wise average of its valid word vectors, where "valid" means that the word is in *Dict*. A linear SVM is trained and we obtain an average accuracy of 67.21% on our dataset. Although this classification result is worse than TF-IDF (82.06%), it can be enhanced by more advanced pooling strategies, rather than a simple average vector over all words, as reported in (Le and Mikolov 2014). Additionally, recall that our data source is the Google search results according to a category name: this step can also reinforce the superiority for word frequency based methods like TF-IDF. On the other hand, since the word vector tries to learn a semantic representation of words with much less dimension, the simple word frequency statistical information will surely lose a lot. However, by late fusion with TF-IDF, we get the score of 84.19%, improving by 2% the single TF-IDF performance, as shown in Table 3.5. TF-IDF and word2vect encode complementary information in textual data.

TF-IDF	word2vec	TF-IDF+word2vec
82.06%	67.21%	84.19%

Table 3.5: Average accuracy of late fusion of TF-IDF and averaged word2vec representations.

The embedded word vector space allows to explore semantic relationships. To investigate this aspect, we report in Table 3.6 the closest words by using the cosine distance metric for *ravioli*, *sushi*, *pho* in the embedded vector space (using the $Dict_v$ dataset). The five most closest words are strongly semantically related to the given query. Additionally, calculating a simple average of the words in a phrase also results in a reasonable semantic. In Table 3.7, we show the closest words of *rice*, *japan* and *rice japan*. As we can see, *koshihikari*, which is a popular variety of rice cultivated in Japan, is closest to *rice japan*, meanwhile *koshihikari* is out of their first five candidates for either *rice* or *japan*. This result signifies that word vector has expressed the semantic of the short phrase *rice japan*. Moreover, *koshihikari* is not among the 101 food category, its meaning and relation with other words are all learned from the corpus in a purely unsupervised

manner. Such a powerful semantic understanding property could help search engine understand user-level needs with natural language as input. It is a promising tool for filling the semantic gap.

ravioli	sushi	pho
gnocchi 0.67	nigiri 0.69	souppho 0.68
tortelli 0.58	maki 0.65	vietnames 0.59
cappellacci 0.55	uramaki 0.65	phos 0.57
delallocom 0.52	sashimi 0.64	beefnoodl 0.58
itemtitlea 0.52	norimaki 0.64	bo 0.56

Table 3.6: 5 most similar words of *ravioli*, *sushi* and *pho* retrieved in the word embedded space. We observe that the identities retrieved are highly semantic relevant.

rice	japan	rice japan
calros 0.59	osaka 0.70	**koshihikari 0.64**
basmati 0.59	tokyo 0.62	awabi 0.61
vermicelli 0.58	kyoto 0.62	japanes 0.61
stirfri 0.58	chugoku 0.61	nishiki 0.59
veget 0.58	gunma 0.60	chahan 0.57

Table 3.7: Short phrase *rice japan* represented by the sum of the word vectors of *rice* and *japan*, is closest to *koshihikari*, which is a kind of japanese rice.

3.5 WEB-BASED RECIPE RETRIEVAL APPLICATION

Providing an efficient way to automatically recognize the food/dish or its recipes on our plates will not only satisfy our curiosity but can have a wider impact on daily life in both the real and virtual worlds. "What is the name of this dish?", "How to cook this?". As a proof of concept for tackling this problem, we create a web search engine [5] that allows any mobile device to send a query image and to get answers to our questions. For any query

5 Available at http://visiir.lip6.fr/

image, the result is a ranking of the 5 best categories automatically found with a matching score. Fig. 3.4 presents the answer to a query image *a cake with strawberry on it*. The category predicted with the highest probability is exactly *Strawberry Shortcake*. The images returned by the application are associated with the hyper-link to the recipe web page.

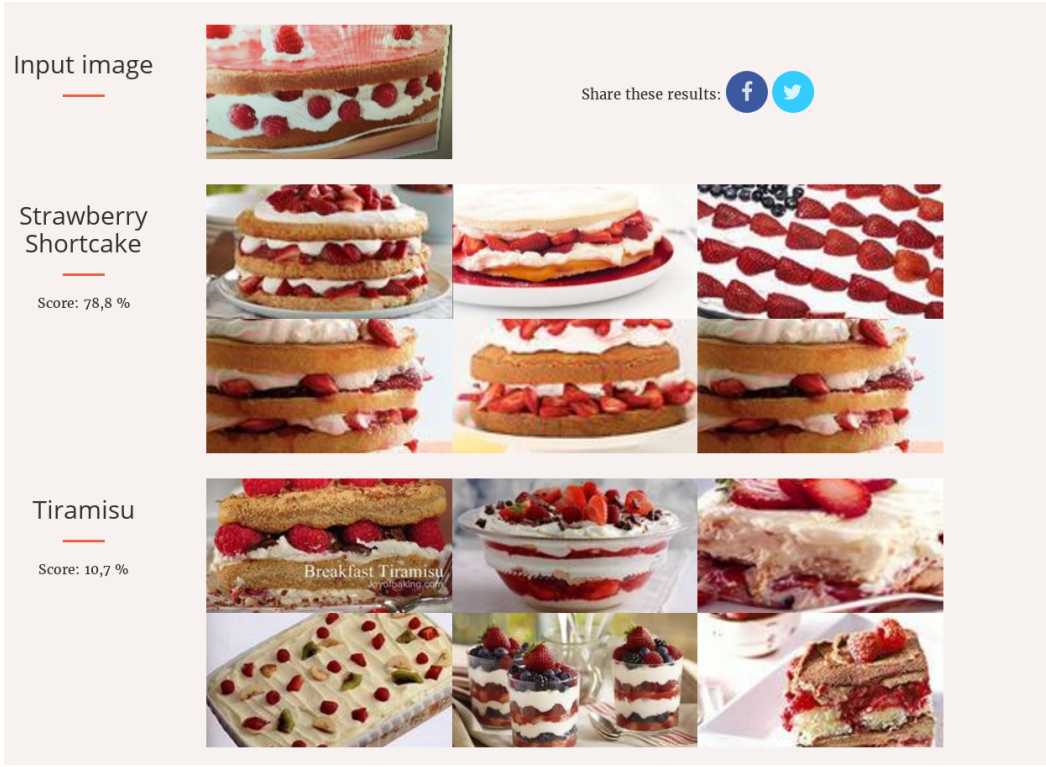


Figure 3.4: The recipe retrieved by our application for a *Strawberry Shortcake* image. (The interface is different from the actual application for better illustration.)

3.6 CONCLUSION

In this chapter, we introduce a large multimedia dataset with 101 food categories. We set the classification baselines for our new dataset by

testing both shallow and deep models. For shallow models, we test the Bag-of-Words Histogram (BoW) + SIFT and Bossanova Image Pooling Representation (Avila et al. 2012). For deep models, we extract directly the features from the models OverFeat and VGG-16 and train an SVM. We also train from scratch and fine-tune the two models on our datasets. As a result, fine-tuning model outperforms others with a significant margin. Our experiments suggest that for visual recognition, fine-tuning deep model is the best step forward. Furthermore, count-based textual features, i.e. TF-IDF, and embedded-based textual features, i.e. Word vector, are coupled with image features for highlighting the complementarity of multimodal data. We find that word vector shows powerful ability in representing any word in a semantical food continuous space. We also run complementary experiments to highlight differences and complementarity of our UPMC Food-101 dataset with the recently published ETHZ Food-101 dataset. Based on our dataset, we propose a retrieval system that we plan to improve using machine learning techniques (Gorisse, Cord, and Precioso 2011; Gosselin and Cord 2008; Picard, Cord, and Revel 2008; Fournier, Cord, and Philipp-Foliguet 2001; Gosselin and Cord 2004; Cord and Gosselin 2006) for user interaction. Part of this chapter has been published in the 2015 IEEE International Conference on Multimedia and Expo as a conference paper, which has been cited by top-level conferences and journals including CVPR, ACM MM, WWW, TMM etc., showing its interest for the scientific community.

4

GAZE BASED WEAKLY SUPERVISED IMAGE CLASSIFICATION

ABSTRACT

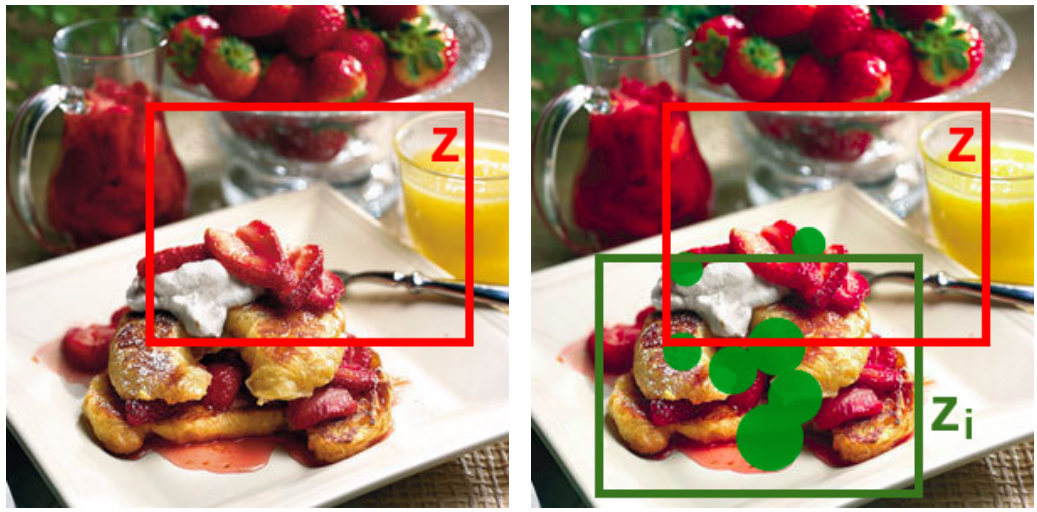
As introduced in previous chapters, human gaze is highly related to the human attention when performing an image classification task. In this chapter, we benefit from this psychological research result and develop a new weak localization model, which leverages human gaze for image classification. For tackling this problem, the commonly used weak supervision is global image label. In our model G+LSVM, intuitively, the region with high gaze density is preferred, while the region with low gaze density is heavily penalized. A gaze density related gaze loss is proposed to compensate the lack of localization information in global image label. The gaze loss and classification loss are jointly optimized as a concave-convex upper bound of the non-convex problem and solved by the Concave-Convex Procedure (CCCP). An appealing feature of G+LSVM is that the model only uses the gazes for training, whereas only visual information is used for prediction. Experimental results show that G+LSVM significantly outperforms LSVM on classification and localization tasks, and that the model achieves similar performance as a model trained with expensive bounding box annotations. Qualitative results also show that the region selected by our model is more semantic meaningful than the baseline.

This work in this chapter is partly published as:

- Xin Wang, Nicolas Thome, and Matthieu Cord (2016). "Gaze latent support vector machine for image classification." In: *IEEE International Conference on Image Processing (ICIP)*, pp. 236–240

As introduced in previous chapters, weakly supervised learning (WSL) and human gaze are both promising for reducing the image labeling cost. For benefiting both the goodness of WSL and human gaze information, recently, attempts have been devoted to incorporating gazes as weak supervision signals (Fathi, Li, and Rehg 2012; Papadopoulos et al. 2014; Ge et al. 2015) for improving the performance of classification or segmentation systems. In (Papadopoulos et al. 2014), objects detectors are trained from gaze features instead of accurate bounding boxes, showing promising results.. In the section 2.3.3 and section 2.4 we introduce eye-tracking and weakly supervised learning in computer vision. In this chapter, we propose a new model, G(aze)+LSVM, which attempts at integrating gaze feature for image classification improved by weakly supervised region selection. Our model generalizes the baseline latent Support Vector Machine (LSVM) (Felzenszwalb et al. 2010) by preferring high gaze density region for localizing objects.

In Fig. 4.1a, when LSVM converges to a bad local minimal, it will predict an inappropriate region as Z for this image *French Toast*. To improve the quality of the region selection, G+LSVM also supports regions with high density of gazes with respect to the region with the highest density of gazes (region Z_i in Fig. 4.1b), by assuming that gaze features are related to regions relevant for the recognition task. For example, if our model still predicts the region Z in Fig. 4.1b for *French Toast*, it will be penalized by a large gaze loss to update the model parameters. Unlike (Shcherbatyi, Bulling, and Fritz 2015; Fathi, Li, and Rehg 2012), G+LSVM only exploits gazes during training phase, and uses only the visual information at test time without gazes. Mathe et al. (Mathe and Sminchisescu 2014; Mathe, Pirinen, and Sminchisescu 2016) proposes using reinforcement learning to find a latent space sampling policy from gaze. This method is efficient at the cost of prediction accuracy. Karthikeyan et al. (Karthikeyan et al. 2013) proposes to train a face and text detector from only gaze information. Although this work does not use image features, it still requires bounding boxes to segment out face and text regions. Shcherbatyi et al. (Shcherbatyi,



(a) LSVM (b) G(aze)+LSVM

Figure 4.1: Gazes bias the selection of latent regions for LSVM. The interpretation is in the section 4.1.

Bulling, and Fritz 2015) integrates gaze into Deformable Part Model for selecting one relevant object location. Their model require gaze annotations for test.

This chapter is organized as follows: The model and optimization procedures of G-LSVM are described in the section 4.2. Experimental results and analysis are shown in the section 4.3. The chapter conclusion is in the section 4.4.

4.2 GAZE-BASED WSL MODEL: G+LSVM

4.2.1 *Latent SVM for image recognition*

We consider the problem of learning from weak supervision in a binary classification context based on the Latent SVM model (Felzenszwalb et al. 2010). The prediction function $f : \mathcal{X} \rightarrow \mathcal{Y}$ takes as input an image x, and

outputs a binary $y \in \{+1, -1\}$. Each image x is associated with latent variables $z \in Z(x)$, which corresponds to a set of sub-regions. For each region z in image x, we extract a visual feature vector $\Phi(x, z) \in \mathbb{R}^d$, *e.g.* deep features. Our model is linear with respect to Φ, *i.e.* each region z is assigned the score $\langle \mathbf{w}, \Phi(x, z) \rangle$, where $\mathbf{w}$ is learned from data. The problem is weakly supervised since the region-specific labels are unknown during training. Our prediction takes the maximum score over the latent variables:

$$f_{\mathbf{w}}(x) = \max_{z \in Z(x)} \langle \mathbf{w}, \Phi(x, z) \rangle. \tag{4.1}$$

A standard classification metric is the 0/1 loss, which means the loss equals 0/1 if the classification is correct/false. However, 0/1 loss is difficult to optimize. As in LSVM, we use the hinge loss as a conventional upper-bound of 0/1 loss. As a result, a classical-SVM like loss is proposed for LSVM:

$$\mathcal{L}_{LSVM}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \Delta_{\text{hinge}}(\hat{y}_i, y_i^*), \tag{4.2}$$

where y_i^* is the true label of image x_i, $\hat{y}_i = \text{sgn}\,(f_{\mathbf{w}}(x_i))$ is the label predicted by our model, hinge loss is defined as $\Delta_{\text{hinge}}(\hat{y}_i, y_i^*) = \max(0, 1 - y_i^* f_{\mathbf{w}}(x_i))$ and $\frac{1}{2}\|\mathbf{w}\|^2$ is the standard max margin regularization term.

4.2.2 *G+LSVM Training*

The novelty of our model is that G+LSVM generalizes latent SVM by biasing the selection of latent regions based on the gaze information during the training scheme. The training objective of G+LSVM is as follows:

$$\mathcal{L}_{G^+}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \Delta_{hinge}(\hat{y}_i, y_i^*) + \gamma \cdot \delta_g(\hat{z}_i, x_i, y_i^*), \tag{4.3}$$

where $\hat{z}_i = \underset{z \in Z(x_i)}{\text{argmax}} \langle \mathbf{w}, \Phi(x_i, z) \rangle$ interpreted as the *relevant region* selected by our model. For each training example, Eq. (4.3) includes a classification hinge loss and a gaze loss δ_g, with a scalar trade-off parameter $\gamma \geq 0$.

In our training scheme, the gaze loss δ_g defined as:

$$\delta_g(\hat{z}_i, x_i, y_i^*) = \begin{cases} 1 - \frac{g(x_i,\hat{z}_i)}{g(x_i,z_i^*)} & \text{if } y_i^* = 1 \\ 0 & \text{if } y_i^* = -1, \end{cases} \tag{4.4}$$

where $g(x_i, z)$ is the total duration of fixations in the region z for image x_i (it can also be seen as the density of gaze for this region), z_i^* is the region which contains the maximum total duration of fixations among all the regions. We only consider the positive image because the gazes in the positive image indicate where the object is. We propose to make use of gaze in the negative image in the next chapter. Fig. 4.2 illustrates the proposed gaze loss. In this example, when the color of heatmap is closer to red, the total duration of gaze is higher. The region containing the maximum total duration of gaze is shown as z_i^* (shown as the green rectangle). The gaze loss of z_i^* is thus defined as 0. The red region z_1 contains a smaller total duration of gaze with respect to the blue region z_2, leading to a larger gaze loss.

The intuition of training G+LSVM is straightforward. Our training objective in Eq. (4.3) is biased by the gaze loss δ_g, so that G+LSVM learns a different model parameter $\mathbf{w}$, which tends to minimize gaze loss compared to LSVM. The final decision of our model is to learn a unique $\mathbf{w}$ by compromising between classification loss and gaze loss. In other words, G+LSVM tries to solve the task of classification and localization simultaneously, thus the *relevant region* is presumed to contain the object of interest, which leads to a better classification result.

4.2.3 *G+LSVM Optimization*

To minimize our training objective function Eq. 4.3, we first show that it can be rewritten as a difference of convex functions, *i.e.*

1) *Classification loss part:* For negative example, $y_i^* = -1$. The second term $1 - y_i^* f_{\mathbf{w}}(x_i)$ in its classification loss is convex because it is a sum of a constant and a maximum over a set of convex functions. As a result, the sum of the classification loss of all negative examples are convex. For positive example, since $y_i^* = 1$, it is not convex. We propose to optimize by

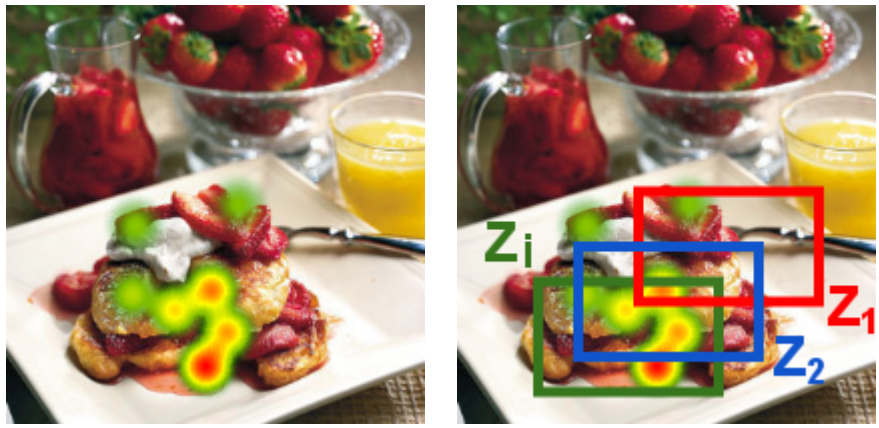


Figure 4.2: The rationale of the definition of gaze loss. When the color of heatmap is closer to red, the total duration of gaze is higher. The region contains the maximum total duration of gaze is shown as z_i^* (shown as the green rectangle). The gaze loss of z_i^* is thus defined as 0. The red region z_1 contains a smaller total duration of gaze with respect to the blue region z_2, leading to a larger gaze loss.

decomposing the hinge loss of positive example into a difference of two convex functions by applying the following theorem:

$$\max(0, u - v) = \max(u, v) - v, \tag{4.5}$$

where u, v are two convex functions. The non-convex classification loss of every positive example is thus decomposed as:

$$\max(0, 1 - f_{\mathbf{w}}(x)) = \max(1, f_{\mathbf{w}}(x)) - f_{\mathbf{w}}(x). \tag{4.6}$$

The maximum of a set of linear functions is convex, so Eq. 4.6 is a difference of two convex functions.

2) *Gaze loss part:* $\delta_g(\hat{z}_i, x_i, y_i^*)$ is difficult to optimize, because the dependency on w is complex and non-smooth. To overcome this issue, we derive

a convex upper-bound Δ_g, inspired from *margin-rescaling* (Joachims, Finley, and Yu 2009):

$$\begin{aligned} \delta_g(\hat{z}, x_i, y_i^*) \leq & \quad \delta_g(\hat{z}, x_i, y_i^*) + \mathbf{w} \cdot \Phi(x_i, \hat{z}) - \mathbf{w} \cdot \Phi(x_i, z_i^*) \\ \leq & \quad \max_{z \in Z(x_i)} [\delta_g(z, x_i, y_i^*) + \mathbf{w} \cdot \Phi(x_i, z)] - \mathbf{w} \cdot \Phi(x_i, z_i^*) \\ := & \quad \Delta_g(\hat{z}, x_i, y_i^*) \end{aligned} \tag{4.7}$$

where $\max_{z \in Z(x_i)} [\delta_g(z, x_i, y_i^*) + \mathbf{w} \cdot \Phi(x_i, z)]$ is a max over linear functions, so it is convex. The second term $\mathbf{w} \cdot \Phi(x_i, z_i^*)$ is linear. As a result, the difference of the two terms is convex.

Aggregating Eq. 4.6 and Eq. 4.7 together, the concave-convex upper bound of the objective function of top G+LSVM is Eq. 4.8:

$$\begin{aligned} \mathcal{L}_{G^+}(\mathbf{w}) \leq \overline{\mathcal{L}_{G^+}(\mathbf{w})} = & \quad \frac{1}{2}\|\mathbf{w}\|^2 + C \Bigg[\underbrace{\frac{1}{n_n} \sum_{i_n=1}^{n_n} \max(0, 1 + f_{\mathbf{w}}(x_{i_n}))}_{cn(\mathbf{w})} + \\ & \underbrace{\frac{1}{n_p} \sum_{i_p=1}^{n_p} \max(1, f_{\mathbf{w}}(x_{i_p}))}_{cp_1(\mathbf{w})} - \underbrace{\frac{1}{n_p} \sum_{i_p=1}^{n_p} f_w(x_{i_p})}_{cp_2(\mathbf{w})} + \\ & \underbrace{\sum_{i=1}^{n} \left([\![y_i = 1]\!] \frac{\gamma_+}{n_p} + [\![y_i = -1]\!] \frac{\gamma_-}{n_n} \right) \cdot \Delta_g(\hat{z}, x_i, y_i^*)}_{g(\mathbf{w})} \Bigg] \end{aligned} \tag{4.8}$$

where n_p, n_n are respectively number of positive examples and negative examples. The coefficient of the gaze loss of negative example γ_- is set to be 0.

For brevity, we rewrite Eq. 4.8 as $u(\mathbf{w}) - v(\mathbf{w})$, where:

$$u(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C(cp_1(\mathbf{w}) + cn(\mathbf{w}) + g(\mathbf{w})). \tag{4.9}$$

$$v(\mathbf{w}) = Ccp_2(\mathbf{w}). \tag{4.10}$$

We then optimize $u(\mathbf{w}) - v(\mathbf{w})$ by CCCP (algo.1). The CCCP algorithm is guaranteed to decrease the objective function at every iteration and to converge to a local minimum or saddle point (Yuille and Rangarajan 2001). In Algo 1, the line 3 involves linearizing the concave part $-v(\mathbf{w})$. We calculate the supergradient $\mathbf{v}_t$ of $-v(\mathbf{w})$ at the point $\mathbf{w}_t$, where $\mathbf{v}_t = -\sum_{i_p=1}^{n_p} \Phi(x_i, \hat{z}_i)$. At line 4, the problem becomes convex, we can use any convex optimization tool for solving this problem.

Algorithm 1: Concave-Convex Procedure

Output: $\mathbf{w}^*$

1 Set $t = 0$, stopping criterion ϵ and initialize $\mathbf{w}$ by $\mathbf{w}_0$, $u(\mathbf{w})$ and $v(\mathbf{w})$ are defined as Eq. 4.9 and Eq. 4.10.

2 **repeat**

3 Find hyperplane $\mathbf{v}_t$ to linearize $-v(\mathbf{w})$: $-v(\mathbf{w}) \leq -v(\mathbf{w}_t) + (\mathbf{w} - \mathbf{w}_t) \cdot \mathbf{v}_t$,

4 Solve $\mathbf{w}_{t+1} = \text{argmin}_{\mathbf{w}} u(\mathbf{w}) + \mathbf{w} \cdot \mathbf{v}_t$,

5 Set t = t+1,

6 **until** $[u(\mathbf{w}_t) - v(\mathbf{w}_t)] - [u(\mathbf{w}_{t-1}) - v(\mathbf{w}_{t-1}))] < \epsilon$;

The step 4 of Algo. 1 requires optimizing w. In this work we use SGD.

$$\Delta L_G(w, i) = w + h_c(w, x_i, y_i^*) + \gamma \cdot h_g(w, x_i, y_i^*)$$

where $h_c(w, x_i, y_i) =$

$$\begin{cases} 0 & \text{if } y_i = -1,\ y_i f_w(x_i) \geq 1 \\ \Phi(x_i, \hat{z}) & \text{if } y_i = -1,\ y_i f_w(x_i) < 1 \\ \Phi(x_i, \hat{z}) - \Phi(x_i, \hat{z}_{t-1}) & \text{if } y_i = 1,\ y_i f_w(x_i) \geq 1 \\ -\Phi(x_i, \hat{z}_{t-1}) & \text{if } y_i = 1,\ y_i f_w(x_i) < 1 \end{cases} \tag{4.11}$$

$$h_g(w, x_i, y_i) = \Phi(x_i, \widetilde{z}) - \Phi(x_i, z_i^*) \tag{4.12}$$

$\hat{z} = \text{argmax}_z w \cdot \Phi(x_i, z)$,
$\hat{z}_{t-1} = \text{argmax}_z w_{t-1} \cdot \Phi(x_i, z)$,
$\widetilde{z} = \text{argmax}_z [\text{gl}(z, z_i^*, y_i) + w \cdot \Phi(x_i, z)]$

w is updeted by $w := w - \alpha_t \Delta L_G(w, i)$, where α_t is learning rate.

Note that given a model parameter $\mathbf{w}$, the *relevant region* $\hat{z}$ only depends on image feature as LSVM, without any gaze information (Eq. 4.1). The benefit of this modeling strategy is that G+LSVM only uses *gaze loss* for training, not for the test. This idea is inspired from *learning using Privileged Information (LUPI)* (Vapnik and Izmailov 2015). The problem addressed by LUPI is that the privileged information is available only at the training stage and is not available at the test stage. By including privileged information into training we obtain a better model, which commits lower generalization error thanks to the localization information for human gaze. This modeling strategy is also practical because models trained with gaze can be applied without gaze annotations.

4.3 EXPERIMENTAL RESULTS

Based on PASCAL VOC 2012 object dataset, Papadopoulos *et al.* (Papadopoulos et al. 2014) annotate 10 object categories with gazes: *aeroplane, cat, dog, bicycle, motorbike, boat, horse, cow, diningtable, sofa*. Different as (Mathe and Sminchisescu 2013), during annotation stage, each of 5 observers is assigned *a specific object* to find.

4.3.1 *Image Datasets*

We validate our ideas on three datasets, PASCAL VOC Action dataset annotated with gaze (short for Action) (Mathe and Sminchisescu 2013), PASCAL VOC Object dataset annotated with gaze (short for POET) (Papadopoulos et al. 2014). Both datasets are collected gaze annotations in task-driven manners. Action contains 4588 images, covering 10 categories. POET contains 6131 images, covering 10 categories out of 20 categories of PASCAL VOC Object dataset. The origin of these images is the `train+val` split of PASCAL VOC dataset. Two sample images of POET and Action are shown with gaze annotations in Fig. 4.3.

In order to compare with the state-of-the-art methods, we follow the standard split of `train`, `val`, `test` set as indicated in PASCAL VOC 2012

Figure 4.3: Gaze annotations. *left*: sample image of POET dataset, *right*: sample image of Action dataset. Different colors indicate different observers.

development kit (Everingham et al. 2015). Since POET contains only 10 out of 20 categories of Pascal VOC 2012 Object, we add back the images of the absent categories in the `train+val` set for training, without gaze information. Finally, our model can be evaluated following the standard protocol. For Action, since by default standard test set requires to identify every person in an image with a bounding box, we conventionally train our model on the training set and test on the validation set. Except for the comparison with the state-of-the-art methods, our experiments are performed by 5 random folds test on the `train+val` set of POET, Action.

4.3.1.1 *Gaze annotations*

The gaze annotations over these datasets are all collected in task-control manners with slight variations.

1. POET uses the *category specific protocol*, which means that each subject has a specific category of object, *e.g.* cat, to look at. Images in POET may have multiple categories. These multiple classes images are annotated with more than one set of annotations. In out tests, for a positive image, we use the corresponding set of annotations, for a negative image, we calculate the fixation duration for each region of each category, then take the maximum fixation duration across the categories as the fixation duration of this region.

2. Action uses the *category group protocol*, which means the subject is required to find a specific group of categories, *i.e.* actions or context. In other words, if a subject is required to find actions, the subject should find all possible actions in the image. The setting of Action is weaker than POET because annotations are only related with a person, not a specific action.

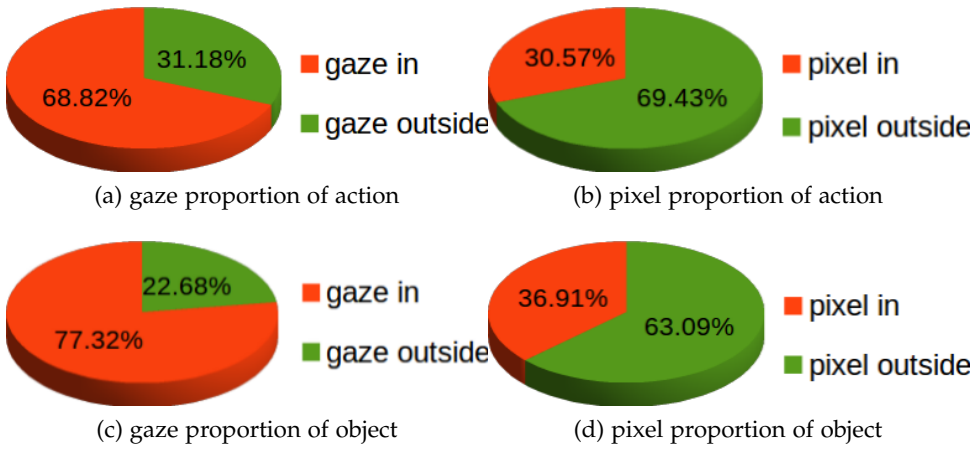
Each gaze is classified into fixation, saccade, or unclassified gaze. For Action and POET, the classification results are already given in the dataset. Gaze is then represented by fixation in the form of a triplet $(x, y, duration)$. (x, y) is the coordinate of fixation, $duration$ is the duration time of this fixation. *Fixation duration* is important since higher exposure time of a fixation reflects a deeper understanding of the local content of the image (Fei-Fei et al. 2007). The total valid fixation time duration of each subject on each image is normalized to a fixed value. By considering the gaze consistency across subjects, for each region, the fixation duration is summed for all subjects. Gaze loss is calculated for each region using the re-weighted summed fixation.

4.3.2 *Statistical consistency of gaze information*

Before evaluating G-LSVM, we first provide a detailed analysis of the gaze data consistency. We compute statistics for the proportion of gazes falling into or outside the bounding boxes of object and compare it to the proportion of image pixels (Fig. 4.4). Statistically, for action dataset, 68.8% of the gazes fall into the ground-truth bounding-box, while the score of pixels is only 30.6%. Similarly, the scores of object dataset is 77.3% *vs* 36.9%. This preliminary study provides a quantitative validation that human gazes are highly related to object localization, and convey relevant features for classification.

4.3.3 *Weakly supervised classification setting*

In our models, the first step is generating the latent regions. Latent region set corresponds to square image regions extracted with a multi-scale sliding



(a) gaze proportion of action

(b) pixel proportion of action

(c) gaze proportion of object

(d) pixel proportion of object

Figure 4.4: Proportions of gazes and pixel numbers in (outside) the ground-truth bounding boxes.

window strategy. Region size vary from 90% to 30% of the whole image area. For a given scale, a window slides from the upper-left to the bottom-right of the image with a step size 10% in both directions. As a result, for each image, the size of sub-region space varies among {4, 9, 16, 25, 36, 49, 64}. Each region is described by the deep features extracted from the FC7 layer of the pre-trained imagenet-vgg-m-2048 deep model[1], which are subsequently L2-normalized and add a bias term. In this setting, the size of feature and model parameter are fixed as 2049.

For training the multi-scale model, we adapt the object bank representation (Li et al. 2010) for our setting. For a given category, we first train the models independently for all 8 scales (including the full image scale). We then form an 8-dimensional vector for each image by the classification scores and train a linear SVM with $C = 10$ as the multi-scale model. Finally, the multi-scale classification score of all categories are averaged to give an mAP to show the overall performance of our models.

1 http://www.vlfeat.org/matconvnet/pretrained/

4.3.4 *Experimental results*

Performance comparison:

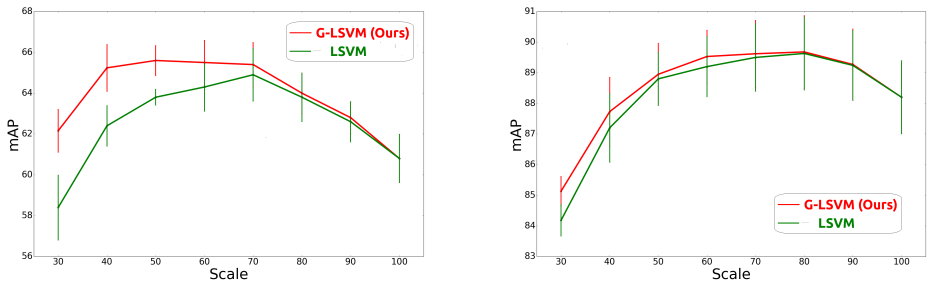
The results are gathered in Table 4.1, using 5 random folds on the `train+val` sets (Everingham et al. 2015), and evaluating performances with the standard mAP metric. We show that G-LSVM outperforms LSVM by a margin of 2.1% for action (resp. 0.4% for object). Paired T-tests reveal that the improvement is statistically significant for a risk of less than 0.5% for action (resp. 2% for object). Both methods largely outperform wSVM, which clearly validate that training WSL models is able to capture local information.

	G-LSVM	LSVM	wSVM
action	70.5 ± 0.8	68.4 ± 1.0	60.8 ± 1.2
object	92.4 ± 1.0	92.0 ± 1.1	88.2 ± 1.2

Table 4.1: mAP(%) of combination multi-scale model.

Fig. 5.5 shows the performance evolution for LSVM and G-LSVM when varying the region scale s. We observe that the improvement of G-LSVM is more pronounced at small scales. This is expected: for large scales, all regions are informative, whereas at smaller scales, the model has to focus on relevant localized features. Note that $s = 100\%$ corresponds to wSVM, for which the mAP for action and object is 60.8% and 88.2% . G-LSVM thus outperforms SVM at all scales of action dataset, as well for scales in $[50, 90]$ on object dataset.

Table 4.2 gives per-class performances at the smallest scale 30%. G-LSVM outperforms LSVM by a margin of 3.1% and 0.8% for respectively action dataset and object dataset. Paired T-tests show that G-LSVM is significant than LSVM for a risk less than 0.5% and 1% for action and object datasets. For action dataset, the performance gain of G-LSVM is especially large for the categories *phoning*, *reading*, *walking*. We note that these actions are usually associated with tiny objects, *i.e.* cellphone, book and small person (*e.g* Fig. 5.8b). For object dataset, G-LSVM performs well at *cow* and *motorbike* and improves over LSVM for most categories.



(a) action dataset (b) object dataset

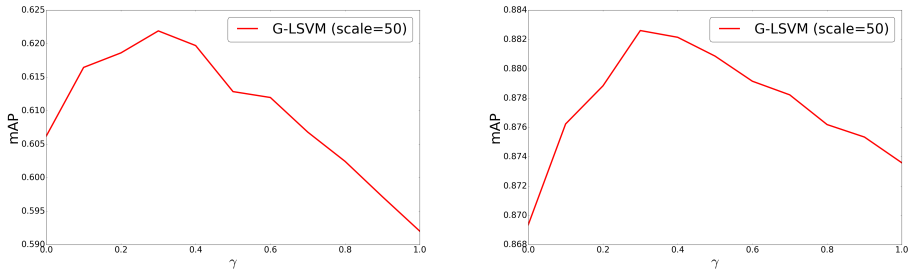
Figure 4.5: mAP(%) at different scales.

Action Dataset	mAP	jump	phone	instru'	read	bike	horse	run	photo	comp'	walk
G-LSVM	**61.29**	**69.20**	**50.51**	**79.49**	**50.57**	78.86	83.88	53.62	**38.64**	72.08	**36.03**
LSVM-Standard	58.17	68.93	41.95	79.21	39.11	**79.26**	**84.20**	**55.11**	36.74	**73.69**	23.52
Object Dataset	mAP	aeroplane	cow	dog	cat	motor	boat	horse	sofa	din'table	bike
G-LSVM	**85.39**	**96.76**	**76.78**	**91.71**	**90.77**	**88.15**	**88.08**	82.82	71.14	**82.08**	**85.59**
LSVM	84.59	96.72	71.97	91.27	90.03	86.30	87.84	**84.05**	**71.19**	81.83	84.75

Table 4.2: AP(%) at scale 30%

Further analysis: The impact of the parameter γ in Eq. (4.3) is shown in Fig. 4.6 for scale 50%. We can see that performances of LSVM, corresponding to $\gamma = 0$, can be improved for most values in $\gamma \in]0, 1.0]$. It is worth noticing that the performances in Fig. 4.6 are shown on average for all classes. We can further substantially boost the performances by cross-validating γ. For example, on the action dataset, a class-wise cross validation ($\gamma \in [0, 1; 0.1]$) at scale 50% leads to nearly 1% improvement compared to $\gamma = 0.2$.

We show in Fig. 4.7 the predicted regions for G-LSVM and LSVM. Results for training images are shown on the first row: we show that G-LSVM selects areas with more gaze features than LSVM. On the second row, we present results for test images, for which gaze features are not available. Interestingly, we can see that G-LSVM extracts regions which are more semantic than LSVM for the classification task.



(a) action dataset (b) object dataset

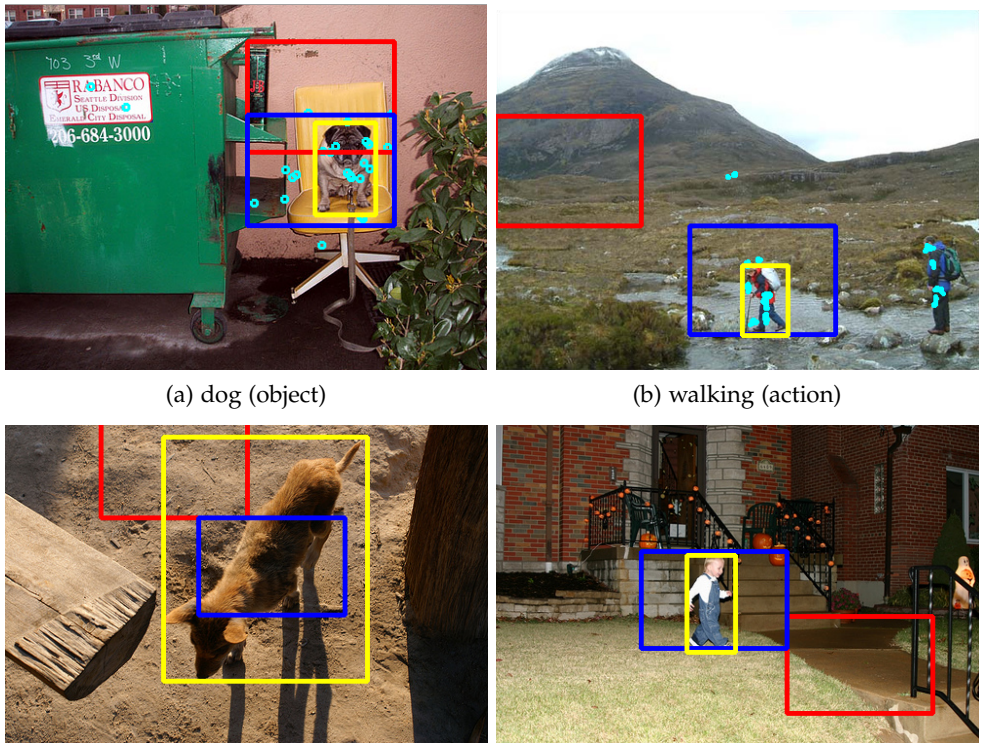
Figure 4.6: For scale = 50%, the effect of parameter γ.

We validate this idea by measuring the detection performances of G-LSVM *vs* LSVM by computing the Intersection over Union (IoU) metric between the predicted region and the ground-truth bounding boxes. The results in Table 4.3 at every scale show that G-LSVM always outperforms LSVM.

action	30	40	50	60	70	80	90
G-LSVM	**21.4**	**25.8**	**27.6**	**28.3**	**29.0**	**29.3**	**28.1**
LSVM	14.5	20.4	24.3	26.7	27.9	28.9	28.0
object	30	40	50	60	70	80	90
G-LSVM	**22.4**	**29.4**	**34.0**	**37.1**	**40.1**	**41.8**	**42.2**
LSVM	20.1	27.1	32.6	36.4	39.2	41.5	42.0

Table 4.3: IoU (%) between predicted region and ground-truth bounding boxes.

Finally, we perform the last experiment using bounding box annotations during training, leading to a model denoted as G-LSVM*. We replace the gaze loss by a ground-truth loss computed as $1 - IoU(z, z_{gt})$, where z_{gt} is the ground-truth region in the dataset. The experiment reveals that G-LSVM is even slightly better than G-LSVM* ($\uparrow$ 0.4 % (0.2%) mAP for the action (object) dataset). This shows that gaze features contain as relevant information as bounding box annotations, while being much cheaper to collect.



(a) dog (object) (b) walking (action)

(c) dog (object) (d) walking (action)

Figure 4.7: Localization results. (a)(b): training results, (c)(d): test results. *red*: LSVM, *blue*: G-LSVM, *yellow*: ground-truth bounding-box. *cyan*: gazes.

4.4 CONCLUSION

In this chapter, we develop a new latent variable model, which leverages human gaze features for image classification using weakly supervised region selection. In our model G+LSVM, a gaze density related gaze loss is proposed to compensate the lack of localization information of global image label. Our model prefers the region with high gaze density, and penalizes the region with low gaze density. The gaze loss and classification loss are jointly optimized as a concave-convex upper bound of the non-convex problem and solved by the Concave-Convex Procedure

(CCCP). An appealing feature of G+LSVM is that the model only uses the gazes for training, whereas only visual information is used for prediction. Experiments show that G+LSVM significantly outperforms the baseline on both PASCAL VOC 2012 object & action classification datasets. We also show that our G+LSVM achieves similar performance when using bounding box annotations, while gaze annotations are much cheaper to collect. Qualitative results show that the region selected by our model is more semantic meaningful than the LSVM baseline.

Although the model outperforms the baseline LSVM in several aspects, it still suffers from two critical deficiencies: 1) it does not exploits the gaze information in the negative images. This is inadequate for data utilization because in the training dataset, the quantity of negative image is usually much larger than the quantity of positive images. 2) on the small scale, selecting a single region is often hard to describe the complete object. In the next chapter, we propose to enhance the G+LSVM by considering the two aspects.

5

MULTI-REGION POSITIVE-NEGATIVE G-LSVM

ABSTRACT

In this chapter, we discuss more deeply the problem of image classification using human gaze in a weakly supervised localization scheme. Based on the the prototype G+LSVM in the previous chapter, we make further improvements in following aspects:

1. We take into account gaze features for negative images whereas only positive images are used in G+LSVM. This is more adequate for data utilization because in the training dataset, the quantity of negative image is usually much larger than the quantity of positive images.

2. We extend the region selection policy from a single region to several regions for performing the prediction, leading to a generalization of top k latent SVM model (Li and Vasconcelos 2015). Selecting more regions not only provides a richer spatial description, but also exploits more extensively the gaze annotation on the image.

We then derive a generalized concave-convex upper bound objective function with respect to the G+LSVM. Also, our model only requires gaze for training, while the test phase is gaze free.

With these improvements, our model k-G$\pm$LSVM generalize better on several image classification benchmark datasets. A thorough experimental analysis validates the proposed model on the standard datasets and our newly proposed gaze food-related dataset UPMC-G20. The UPMC-G20 is annotated with gaze using a task-driven protocol, in where the images are all food images in the UPMC Food-101 dataset. We make this dataset because food is also a complicate object made up of various ingredient and often with several common regional backgrounds. We find our model is capable of capturing more complete semantic regions in both positive

and negative images with respect to the G+LSVM. For example, a *tiramisu* image is often composed of main ingredients. and also backgrounds like forks, plates, strawberry decoration, etc.

This work in this chapter is partly published as:

- Xin Wang, Nicolas Thome, and Matthieu Cord (2017). "Gaze Latent Support Vector Machine for Image Classification Improved by Weakly Supervised Region Selection." In: *Pattern Recognition*, pp. –. DOI: https://doi.org/10.1016/j.patcog.2017.07.001. URL: http://www.sciencedirect.com/science/article/pii/S0031320317302625

To incorporate gaze information into latent SVM model, we proposed in Chapter 4 a new WSL model (G+LSVM) to learn gaze-biased region from image with global label and gaze annotation. As introduced in the section 2.4.1, extending weakly supervised learning model to select more relevant regions (Li and Vasconcelos 2015) and to leverage the information of negative examples (Azizpour et al. 2015; Durand, Thome, and Cord 2015; Durand, Thome, and Cord 2016; Durand et al. 2017) are promising. Related to our work, Shapovalova et al. (Shapovalova et al. 2013) focuses on WSL recognition by penalizing region selection with gaze. However, the gaze information is not sufficiently exploited because only positive examples are penalized with gaze. In this chapter, comparing to the previous works, our model is generalized to leverage the gaze information in both positive and negative examples in multiple regions.

The reason for making these improvements is straightforward: on one hand, using the gaze in the negative images is not the same as for positive image because there is a contradiction between the region selected and the region penalized. The reason is that for an negative image, the selected region is often the background, because positive image and negative image share the background like forks, plates, strawberry decoration in the food images. We discuss in detail the reason of this contradiction and our proposed solution in the subsection 5.2.1. On the other hand, taking only the maximum scored region as the representative is rigid because small-scale region may be to small to fit an object. To soften the constraint, (Li and Vasconcelos 2015) proposes the definition of soft bags of top k instances. In soft bags, example is represented by the average feature of the top k instances. In our model, selecting more regions not only provides a richer spatial description, but also exploits more extensively the gaze annotation on the image. Our final model k-G$\pm$LSVM is presented in the subsection 5.2.2

With these improvements, our model k-G$\pm$LSVM generalizes better on several image classification benchmark datasets. A thorough experimental analysis validates the proposed model on the standard datasets and our

newly proposed gaze food-related dataset UPMC-G20. The UPMC-G20 is annotated with gaze using a task-driven protocol, in where the images are all food images in the UPMC Food-101 dataset. We make this dataset because except for the object or the person action in the PASCAL VOC 2012 can be regarded as a composition of regional parts, food is also a complicate object made up of various ingredient and often with several common regional backgrounds. For example, a *tiramisu* image is often composed of main ingredients. and also backgrounds like forks, plates, strawberry decoration, etc. The detail of UPMC-G20 is described in subsection 5.3. The result shows that our model is capable of capturing more complete semantic regions in both positive and negative images with respect to the G+LSVM across the three datasets.

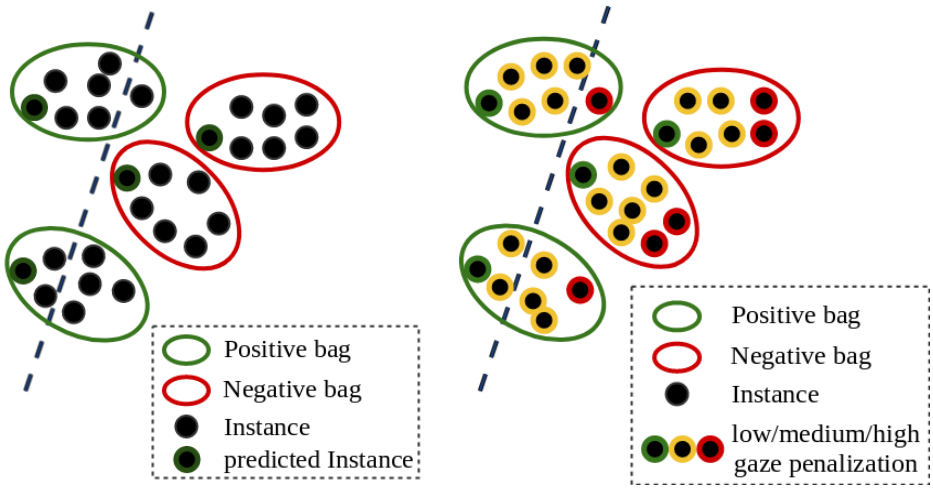
This chapter is organized as follows. In section 5.2 we formally introduce our k-G±LSVM model and the optimization scheme. In section 5.4, we present our experimental results to validate our models. The conclusion is provided in section 5.5.

5.2 k-G±LSVM: WEAKLY SUPERVISED GAZE LATENT SVM

5.2.1 *G±LSVM: Positive Negative Latent SVM*

In the previous chapter 4, we introduce our baseline models LSVM and G+LSVM. One drawback of G+LSVM is the absence of gaze information in negative image. However, a straightforward application of positive gaze loss on the negative image may not work. The reason is that for the positive image, the model should tend to localize where the foreground object is. For the negative image, however, the model should tend to localize where the background is (Azizpour et al. 2015). That's because the overlapping instances between positive and negative example are likely to be the background *e.g.* mint leaves, plates, forks, etc. According to the *task-driven* protocol, image semantic is related with gaze distribution. Indicated by the gaze, the region with lower density of gaze is more likely to be background. Since then, we should penalize the object region of negative image. This intuition is shown in Fig. 5.1b and leads to a generalization of

G+LSVM, called G±LSVM. In G±LSVM we defined a negative gaze loss, which prefers the region where there is less possibility to contain an object. Contrary to positive image, if a region of negative image contains more gaze, it is force not to be the relevant region of the negative image.



(a) LSVM (b) G±LSVM

Figure 5.1: G±LSVM generalizes LSVM by penalizing background of positive image and foreground of negative image.

Based on this assumption, we propose a negative gaze loss defined as follows:

$$\delta_g(\hat{z}_i, x_i, y_i^*) = \begin{cases} 1 - \frac{g(x_i,\hat{z}_i)}{g(x_i,z_i^*)} & \text{if } y_i^* = 1 \\ \frac{g(x,\hat{z}_i) - g(x_i,z_i^{-*})}{g(x_i,z_i^*) - g(x_i,z_i^{-*}) + \epsilon} & \text{if } y_i^* = -1 \end{cases} \tag{5.1}$$

where $g(x_i, z_i^{-*})$ is the minimum number of gaze among all regions of image x_i, ϵ is set to be 10^{-6}. We subtract the term $g(x_i, z_i^{-*})$ from the numerator and denominator only to normalize the minimum negative gaze loss to be 0.

We introduce independent parameters γ_+ and γ_- for trading positive gaze loss and negative gaze loss. Assembling all together we get the objective function of G±LSVM:

$$\mathcal{L}_{G^{\pm}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \Delta_c(\hat{y}_i, y_i^*) + (\llbracket y_i^* = 1 \rrbracket \gamma_+ + \llbracket y_i^* = -1 \rrbracket \gamma_-) \cdot \delta_g(\hat{z}_i, x_i, y_i^*) \tag{5.2}$$

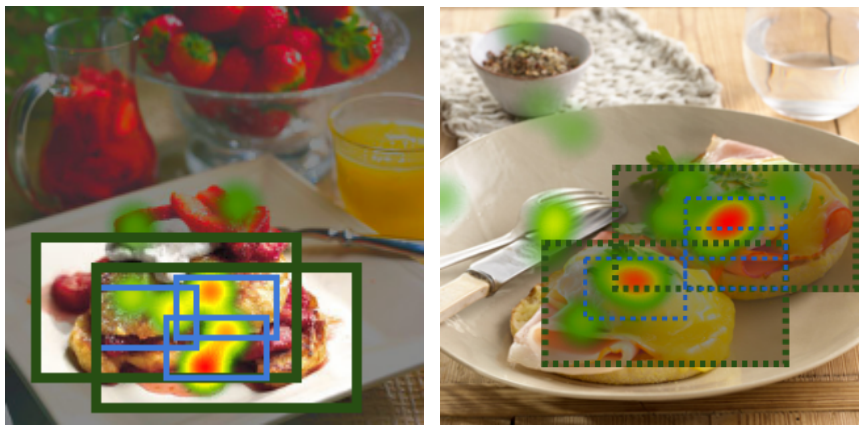
The prediction function follows the LUPI explanation described in section 4.2, so it is also gaze free during the test (same for the following model).

5.2.2 *k-G±LSVM: Top k Positive Negative Latent SVM*

Taking only the maximum scored region as the representative is rigid because one region may be to small to fit an object. To soften the constraint, (Li and Vasconcelos 2015) proposes the definition of soft bags of top k instances. In soft bags, example is represented by the average feature of the top k instances. This method is proved to be robust to the noise in the examples and generalized better than LSVM.

An useful property of top k related to gaze information is its smooth functionality for *sparse gaze limitation*. This limitation is due to the truth that gaze on an image often focus on a small part of the image. For a given example, the gaze loss term has no difference on regions with the same gaze loss. Selection among these regions is random in previous single instance models. This randomness can be eliminated by taking them all via top k strategy.

Fig. 5.2 illustrates the rationale of our final model. Remind that the goal is to select semantically meaningful regions, *e.g.* those containing the target object class (*eggs benedict* region or its sub-regions in Fig. 5.2a). By assuming that gaze features are related to regions relevant for the recognition task, gaze and object are matched for positive example. For negative example, top k G±LSVM further supports regions with low density of gaze, by assuming that no gaze features are related to classify negative images. Extending the model to top k instances latent SVM can further improve the quality of region selection and reduce the effect of the sparseness of gaze.



(a) *french toast*, positive example (b) *eggs benedict*, negative example

Figure 5.2: Illustration of top k-G±LSVM model. Human gaze density is represented by the heat map. In our models, positive example emphasize the latent regions with high gaze density (inside the solid boxes), while negative example emphasizes the regions with low gaze density (outside the dashed boxes). Different colors of regions indicate different scales. For one scale, our model takes multiple highest scored regions as the relevant regions. (Best viewed in color)

The objective function of top k G±LSVM is as follows:

$$\mathcal{L}_{kG^{\pm}_{-}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \Delta_c(\hat{y}_i, y_i^*) + (\llbracket y_i^* = 1 \rrbracket \gamma_+ + \llbracket y_i^* = -1 \rrbracket \gamma_-) \cdot \delta_{\mathbf{g}}(\hat{\mathbf{z}}_i, x_i, y_i^*) \tag{5.3}$$

where

$$\Delta_c(\hat{y}_i, y_i^*) = \max(0, 1 - y_i^* f_{\mathbf{w}}(x_i))$$

$$\delta_{\mathbf{g}}(\hat{\mathbf{z}}_i, x_i, y_i^*) = \frac{1}{k} \sum_{j=1}^{k} \delta_g(\hat{z}_{ij}, x_i, y_i)$$

$$\hat{\mathbf{z}}_i = \underset{\mathbf{z} \in \mathbf{Z}(x_i)}{\operatorname{argmax}} \langle \mathbf{w}, \Phi(x_i, \mathbf{z}) \rangle,$$

where $\mathbf{z}$ is a vector of latent variables, $\mathbf{Z}(x_i)$ the hypothesis space $\{0,1\}^k \setminus \{\mathbf{0}\}$. $\Phi(x_i, \mathbf{z}) = \frac{1}{k}\sum_{j=1}^{k} \Phi(x_i, z_{ij})$.

In the section 5.2 we propose three variations, G±LSVM, top k G+LSVM, top k G±LSVM. Each of the models has a different objective function to optimize. However, notice that when $k = 1$, top k models reduce to the single instance model. Furthermore, when $\gamma_- = 0$, the objective function of G±LSVM (Eq. 5.2) reduces to G+LSVM (eq. 4.3), and when $\gamma_+ = 0$, G+LSVM reduces to LSVM (eq. 4.2). For the reason above, we can refer the optimization of all models to the section 4.2.3. For the model which leverages negative image, the γ_- should be selected, while for the model selecting k top regions, the feature vector and gaze loss are represented as the average of these k instances.

5.3 UPMC-G20 FOOD GAZE DATASET

5.3.1 *UPMC-G20 content*

UPMC-G20 is a food-related gaze annotated dataset based on a multi-modal large scale food dataset UPMC-food 101 (Wang et al. 2015a). We select 20 food categories from UPMC-food 101, resulting in 2,000 images. The images selected do not contain text, because it's verified that texts attract attention most (Wang and Pomplun 2012). For UPMC-G20, I-VT filter (Olsen 2012) is used to classified the gaze into saccade or fixations . For each image, about 15 fixations across 3 subjects (in average) with a total duration of 2.5 seconds are collected. In total, we have collected 31104 fixations. The categories selected are apple-pie, bread-pudding, beef-carpaccio, beet-salad, chocolate-cake, chocolate-mousse, donuts, beignets, eggs-benedict, croque-madame, gnocchi, shrimp-and-grits, grilled-salmon, pork-chop, lasagna, ravioli, pancakes, french-toast, spaghetti-bolognese, pad-thai.

Samples of images and gaze annotations are shown in Table 5.1 and Table 5.2. For full visualization of UPMC-G20, we refer our reader to this page of our dataset: `http://webia.lip6.fr/~wangxin/upmcg20/`.

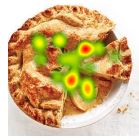
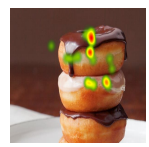
Category	samples			
apple pie				
bread pudding				
beef carpaccio				
beet salad				
chocolate mousse				
chocolate cake				
beignets				
donuts				
croque madame				
eggs benedict				

Table 5.1: Sample image and annotation of UPMC-G20 (1)

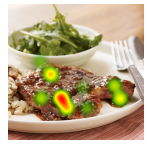
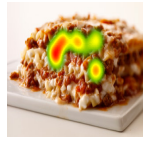
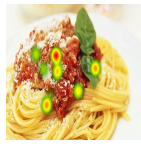
Category	samples			
gnocchi				
shrimp and grits				
grilled salmon				
pork chop				
lasagna				
ravioli				
pancakes				
french toast				
spaghetti bolognese				
pad thai				

Table 5.2: Sample image and annotation of UPMC-G20 (2)

5.3.2 *Apparatus*

Our eye-tracker is a non-invasive Tobii X2-30 with a double eyes gaze sampling rate 30Hz. Eye-tracker is fixed under a 12.6" laptop screen with resolution 1366×768. The subject sits at a distance of about 60cm to the screen. The test environment is quiet and of suitable temperature for not introducing physiological error. The experiment was conducted with the software Tobii Studio (V3.4.5) (AB 2016). Before annotating, for each subject, dominant eye, gender, age are recorded. Before every experiment, Tobii X2-30 is calibrated and validated with a standard nine-point procedure to ensure the coordinate of the gaze recorded matches where the subject is looking at. They are taught the procedure of annotation with a clear explanation and validate a simulation test before the formal experiment. Subject record his classification answer by clicking the corresponding option on the screen after viewing an image using a mouse. Comparing to pressing a button to indicate the category as in (Papadopoulos et al. 2014), using the mouse is useful because mousing moving leads to eye moving after every image. The subject then break the possible steady fixating strategy.

5.3.3 *UPMC-G20 collection protocol*

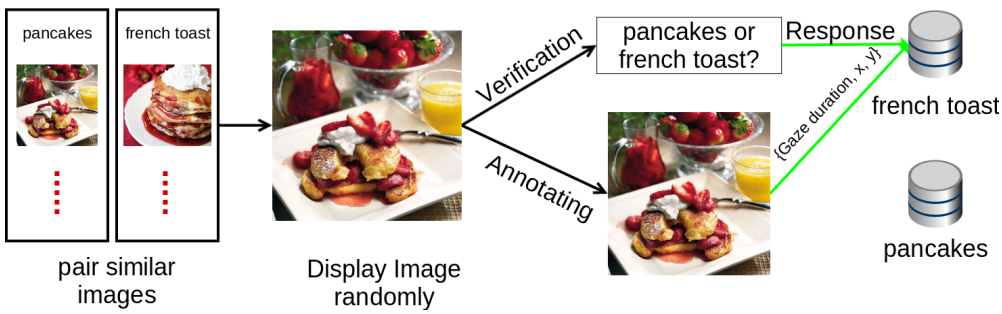


Figure 5.3: Food gaze collection protocol

Our collection protocol is shown in Fig. 5.4. It is inspired by the *two-alternative forced choice object discrimination* (Papadopoulos et al. 2014). This

protocol is simple to the annotators and can save the time because no irrelevant images for distracting the attention are shown.

The protocol is composed of steps:

1. Randomly selecting an image from a pair of categories and exposing for 2.5 seconds, recording the gaze data.

2. Making the subject answering a multiple choice question, of which the category of the image is asked to be selected using a mouse.

3. After exposing every 20 images, a page indicates the progress of the task is shown to heal the anxiety of annotators.

4. After exposing a whole set of images, annotator gets an adequate rest then recalibrate for the next set of images.

5.3.4 *Motivation of constructing the UPMC-G20*

We make the UPMC-G20 dataset because food is a complicated object made up of various ingredients and often with several common backgrounds. In Fig. 5.4, we illustrate our motivation of constructing this dataset by comparing *french toast* and *eggs benedict* images. As we introduce in Chapter 4, it is obvious that the gaze covers the foreground. Our model is able to distinguish the semantic region of food image. And related to the proposition in this chapter, the two images share the background of certain objects like *plate*, *cup*, *fork*. Given *french toast* as the positive example, the prediction of negative example should be the background objects because they have a closer distance to the positive image.. This context is highly related to the application scope of the k-G$\pm$LSVM model. We report detailed experimental results in the next section for supporting our points of view.

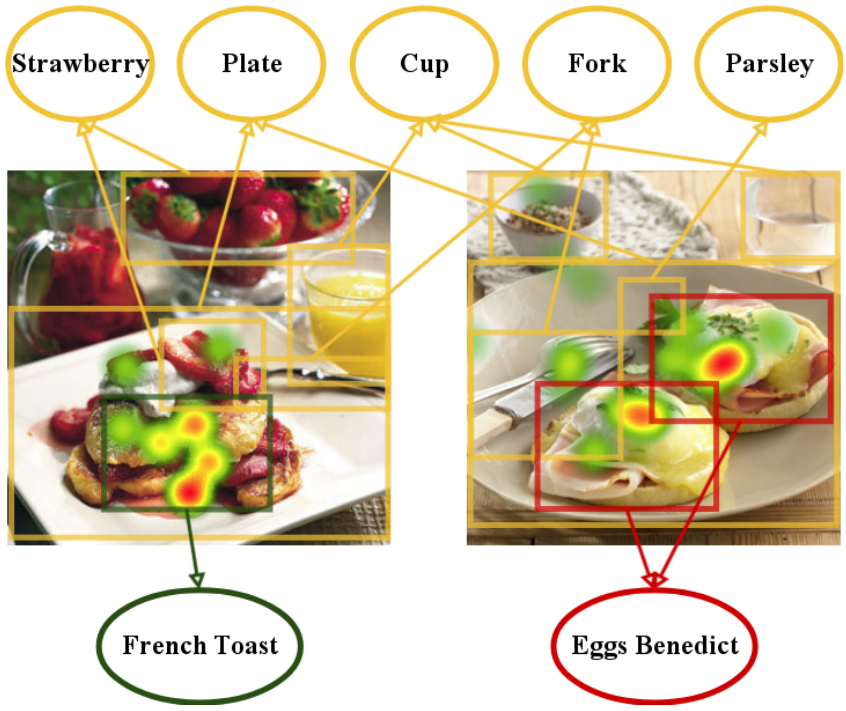


Figure 5.4: Motivation of constructing the UPMC-G20: Food is a complicated object made up of various regional parts and often with several common backgrounds.

5.4 EXPERIMENTS

5.4.1 *Comparison with the state-of-the-art*

In our model, we set k-G±LSVM with the parameters $C = 10^4$, $\gamma_+ = 0.2$, $\gamma_- = 0.05$ for each scale. In a heuristic manner, for top k models, we set $k = 2, 4, 6, 8$ for scale 90% to 60%, and $k = 10$ for scale 50% to 30%. A multi-scale model is trained as indicated in section 4.3.3. In all experiments, we use the standard metric mean Average Precision (mAP) as for PASCAL VOC classification. The basic setting of our experiments is the same as section 4.3.

In Table 5.3 we show the global score of different methods on the three datasets and the annotations they use. For POET dataset, Deep Fishing (Gordo, Gaidon, and Perronnin 2015) and Z&F network (Zeiler and Fergus 2014) are two deep network based methods, which only use image label for training. NUS-SCM (Song et al. 2011) is an SVM-based method and Oquab (Oquab et al. 2014) is a fine-tuned deep network. They both use training bounding box as the additional annotation. Our method outperforms the four methods with only our weak supervision signals. For Action dataset, we compare with Action part (Gkioxari, Girshick, and Malik 2015) and RMP (Hoai 2014). The action part is a deep version of poselets and capture parts of the human body under a distinct set of poses, while RMP considers deformation of discriminative parts. They both propose a model with simple annotations (*e.g.* image label and training bounding box) and a model with rich annotations (*e.g.* test bounding box and part annotation. Our model is better than them if they do not use rich annotations. In Table 5.4 we show the per category performance on the test set of POET. Our model largely outperforms other methods on *boat*, *cat* and *diningtable* categories.

	Action	POET	label	train BB	test BB	part	gaze
Deep Fishing (Gordo, Gaidon, and Perronnin 2015)	-	79.9	√				
Z&F (Zeiler and Fergus 2014)	-	81.2	√				
RMP (Hoai 2014)	65.1	-	√				
NUS-SCM (Song et al. 2011)	-	84.3	√	√			
Oquab (Oquab et al. 2014)	-	84.5	√	√			
Action part (Gkioxari, Girshick, and Malik 2015)	64.6	-	√	√			
RMP (Hoai 2014)	71.4	-	√	√	√		
Action part (Gkioxari, Girshick, and Malik 2015)	71.0	-	√	√	√	√	
k-G±LSVM (ours)	69.6	85.9	√				√
G+LSVM (Wang, Thome, and Cord 2016)	66.8	82.6	√				√
wSVM	59.1	79.8	√				

Table 5.3: Comparison with the state-of-the-art methods on the test set of Pascal VOC 2012 Object, and the validation set of *Action*. Our model outperforms other methods even when they use `global label + training bounding box`. We also achieve comparable results with respect to the models using accurate annotations such as `test bounding box` and/or `human part annotation`.

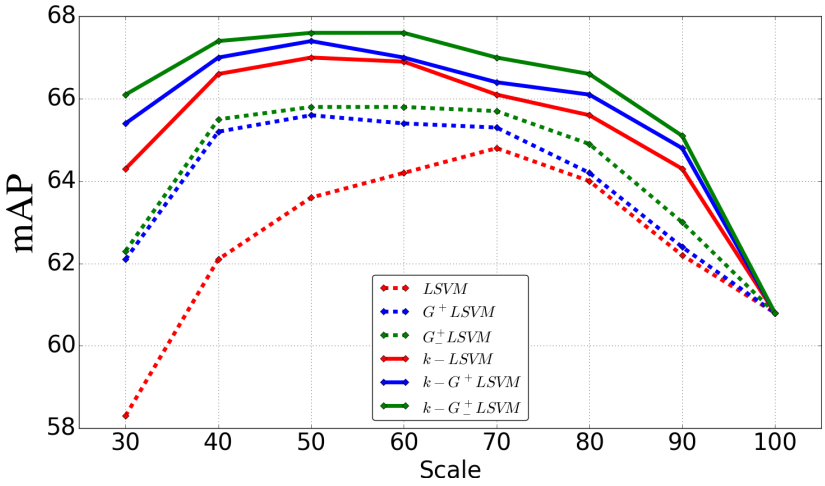
POET	mAP	plane	bike	boat	cat	cow	table	dog	horse	motor	sofa
Deep Fishing (Gordo, Gaidon, and Perronnin 2015)	79.9	95.0	76.6	82.9	88.6	65.4	69.8	86.5	82.1	85.1	57.0
Z&F (Zeiler and Fergus 2014)	81.2	96.0	77.1	85.5	91.2	74.4	67.7	87.8	86.0	85.1	61.1
NUS-SCM (Song et al. 2011)	84.3	**97.3**	**84.2**	85.3	89.3	77.8	75.1	83.0	**87.5**	**90.1**	**73.4**
Oquab (Oquab et al. 2014)	84.5	94.6	82.9	84.1	90.7	**86.8**	69.0	92.1	93.4	88.6	62.3
k-G±LSVM (ours)	**85.9**	97.2	83.9	**90.1**	**94.7**	77.4	**77.3**	**92.3**	87.3	89.9	68.9
G-LSVM (Wang, Thome, and Cord 2016)	82.6	96.5	80.2	87.7	92.4	71.1	74.1	89.6	84.3	87.5	62.7
wSVM	79.8	95.4	79.6	86.7	92.2	59.6	69.9	90.0	86.7	79.3	58.4

Table 5.4: mAP(%) per category on the test set of PASCAL VOC 2012 Object.

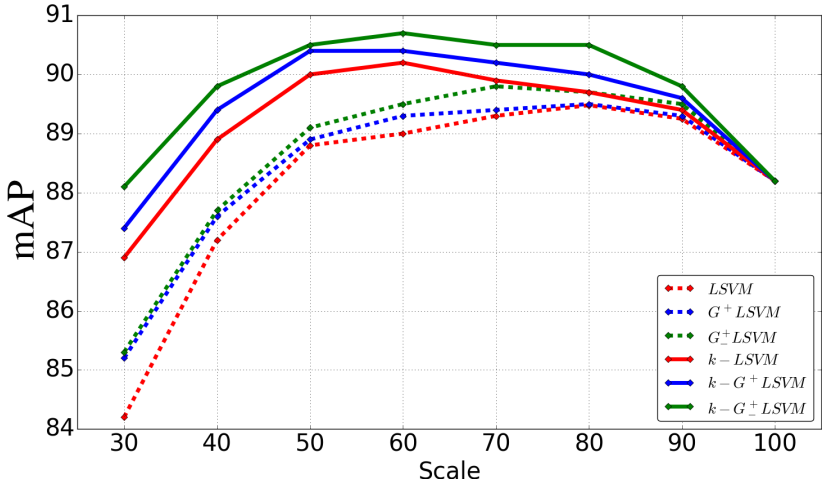
5.4.2 *Ablation studies*

In this section, we compare LSVM, G+LSVM, G±LSVM and their top k variations. We present the scale-wise classification experiments in Fig. 5.5. In our model, scale measures the size of the sliding window with respect to the size of the image. In a heuristic manner, for top k models, we set $k = 2, 4, 6, 8$ for scale $90\% - 60\%$, and $k = 10$ for scale $50\% - 30\%$. For most scales, the model performance is better than wSVM (scale=100 in Fig. 5.5). This result proves the effectiveness of weakly supervised learning: local information is critical for image classification.

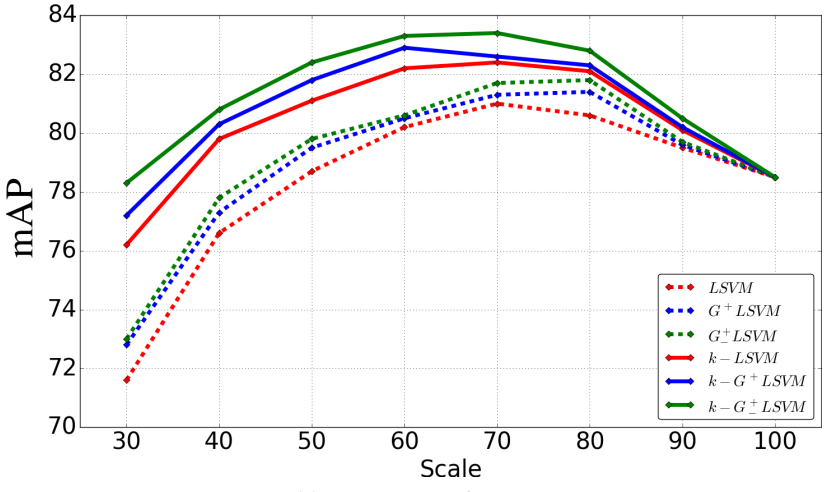
We can also observe that adding gaze into the model improve the performance for all scales. The improvement can be explained by two reasons. One is that G+LSVM emphasizes small scales. That is what we expect: for large scales, nearly all regions of positive images are informative, whereas at smaller scales, the model has to focus on relevant localized features. The other is that G±LSVM can also emphasize large scales. Paired T-tests show that G±LSVM is better than LSVM with a larger significance than for G+LSVM, especially for large scale. This phenomenon may have a dual explanation with respect to G+LSVM: not all regions of negative images are non-informative. As a result, for large scale, the ground truth region z_i of negative example has a larger probability to be unique. While for small scales, z_i is selected randomly among all low gaze density regions, which may lead to a less optimal result. When k increases, for small scale, this problem no longer dominates the performance because the set of ground truth regions for negative images is informative with less randomness.



(a) Action dataset



(b) POET dataset



(c) UPMC-G20 dataset

Figure 5.5: mAP(%) at different scales. In our model, scale measures the size of

We think that is the reason why we observe a substantial performance enhancement at small scales for top G±LSVM.

Table 5.5 gives the performance at the smallest scale 30%. At scale 30%, k-G±LSVM (k-G+LSVM) outperform k-LSVM by a margin of 1.8%(1.1%), 1.2%(0.5%), 2.3%(1.2%) for respectively Action, POET and UPMC-G20. Paired T-tests show that k-G±LSVM (k-G+LSVM) is more significant than LSVM for a risk less than 0.2%(1.0%), 1.0%(2.0%), 0.2%(0.5%) for respectively Action, POET, UPMC-G20. These statistical results show that k-G±LSVM is better than k-G+LSVM with significance at small scale. Top k models much outperform single instance models. Interestingly, as we expected, the gain of k-G±LSVM with respect to k-G+LSVM is much larger than the gain of G±LSVM with respect to G+LSVM.

	Action	POET	UPMC-G20
k-G±LSVM	66.0 ± 0.9	88.1 ± 1.2	78.3 ± 1.0
k-G+LSVM	65.3 ± 1.0	87.4 ± 1.0	77.1 ± 1.1
k-LSVM	64.2 ± 0.8	86.9 ± 1.1	76.0 ± 1.2
G±LSVM	62.4 ± 0.9	85.3 ± 1.1	$73.0. \pm 0.8$
G+LSVM	62.1 ± 0.8	85.2 ± 1.0	72.9 ± 0.9
LSVM	58.2 ± 1.0	84.2 ± 1.1	71.6 ± 1.0

Table 5.5: mAP(%) of scale 30% on Action, POET and UPMC-G20 datasets. Here we set $k = 10$.

5.4.3 *Study of hyper-parameters*

We investigate the impact of the three hyper-parameters in our model: trade-off parameters γ^+, γ^- and k. The impact of the parameter γ_+ of G+LSVM is shown in Fig. 5.6 for small scale 50%, with k set to be 1. The performances in Fig. 5.6 are shown on average for all categories. For all three datasets, mAP reaches the peak when γ_+ is in the interval $[0.1, 0.3]$. Note that when γ_+ gets too high, mAP gets even lower than not adding gaze. Fix γ^- to be the best value obtained by cross-validation, for γ^- , the effective value is found to be a relatively small value between $[0.05, 0.1]$.

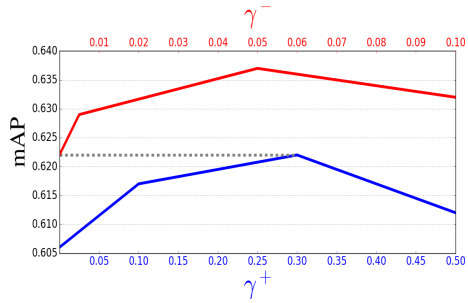
This result is reasonable because our objective is classification with gaze information as auxiliary information, so the gaze loss should tend to have a smaller weight than the classification loss. The performance of k model varies in the similar trend.

We show in Fig. 5.6 that our model outperforms k model significantly for all k value at scale 30%. We set γ_+ of G±LSVM and G+LSVM to 0.2, γ_- of G±LSVM to 0.05. From Fig. 5.6, we also find that by increasing k, gaze latent SVM always outperforms latent SVM. This result signifies that gaze helps better select the regions even when the number of candidate regions largely increases. Heuristically, for selecting k, the small scales prefer a larger k. That's because, for small scale, more regions are semantic for positive images and can smooth the selection of ground-truth regions of negative examples.

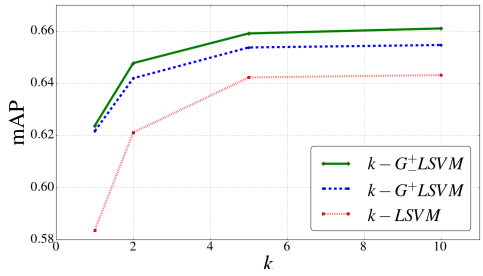
5.4.4 *Localization results*

The relevant regions proposed by our models are interpretable. We show in Fig. 5.7 and 5.8 the predicted regions for the model k-G±LSVM at scale 30%, where $k = 10$. We present the first three high scored regions for visual clarity.

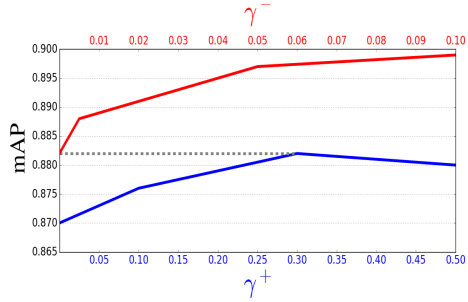
Results for training images are shown in the first row: we show that k-G±LSVM selects areas with more (fewer) gaze for positive (negative) images. Results for test images are shown in the second row, of which gaze features are unavailable. The k-G±LSVM extracts regions which are highly semantic for positive images and extract background for negative images. For example, we find that *running* and *french toast* model has a good result on the positive images. Also for the negative image, the *running* focuses on the regions, which have similar visual semantic to the *road* and *trees*, and the *french toast* model focuses on the regions, which have similar visual semantic to the *cups* and *strawberries*. As these regions have a relatively low density of gaze, our model does not penalize them too much. Interestingly, these regions often appear as the background in the positive images.



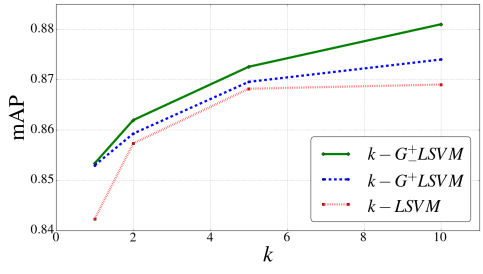
(a) Action dataset



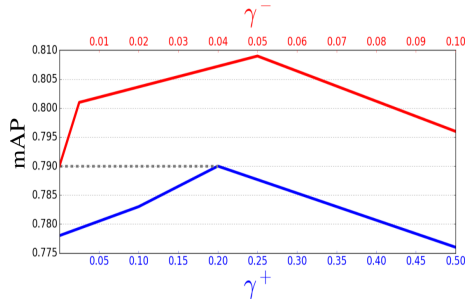
(b) Action dataset



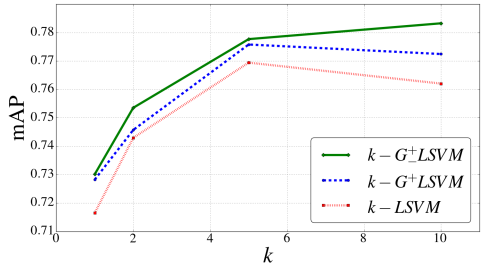
(c) POET dataset



(d) POET dataset

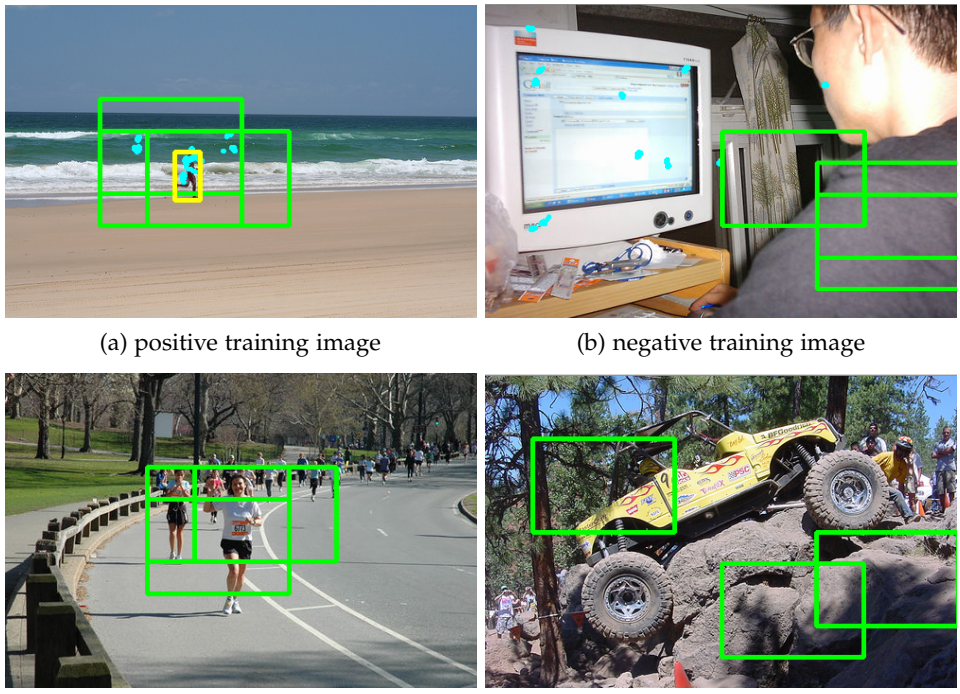


(e) UPMC-G20 dataset



(f) UPMC-G20 dataset

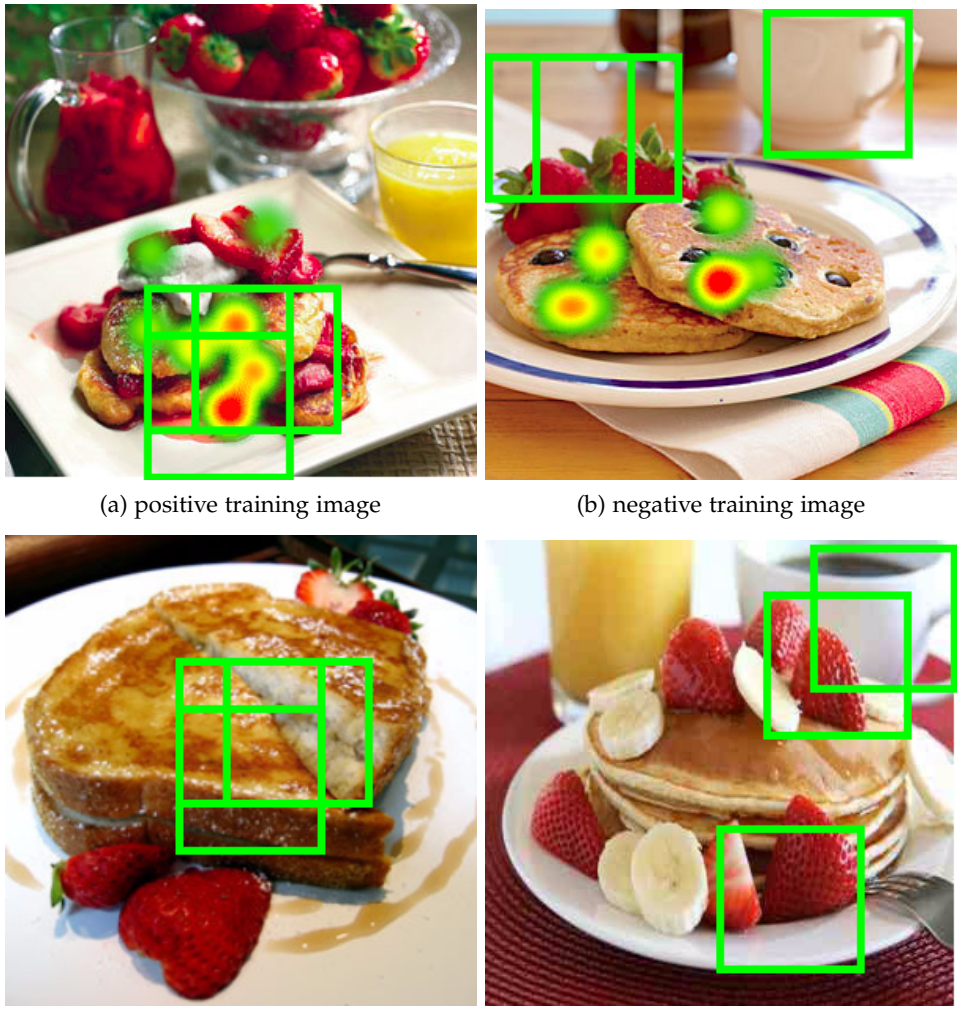
Figure 5.6: The sensitivity of hyper-parameters γ_+ and k. *left*: At scale 50%, the performance with respect to γ_+ (γ_-) is found to reach the peak value in the interval $[0.1, 0.3]$ ($[0.05, 0.1]$). *right*: At scale 30%, generally, the larger k is, the better the performance is.



(a) positive training image (b) negative training image

(c) positive test image (d) negative test image

Figure 5.7: Localization results achieved by *running model*. (a)(b): training results, (c)(d): test results.



(a) positive training image (b) negative training image

(c) positive test image (d) negative test image

Figure 5.8: Localization results achieved by *french toast model*. (a)(b): training results, (c)(d): test results.

5.5 CONCLUSION

In this chapter, the model G±LSVM extends and outperforms our previous model in several aspects. G±LSVM exploits gaze for guiding the selection of regions, which are relevant with the image semantic. We find that generalizing the model to the selection of k maximum scored regions can also benefit from the gaze information. G±LSVM also only leverages human gaze for training, while the test is gaze free. Experimental results show that G±LSVM achieves competitive results with respect to the state-of-the-arts methods on Pascal VOC Action and Object. We also publish a food-related dataset annotated with gaze, UPMC-G20, for further validating the generalization ability of our models. The experimental results show that G±LSVM model also achieve reasonable performance on UPMC-G20. The qualitative analysis shows that the regions selected by our model are highly semantic across the three benchmark datasets.

6

CONCLUSION & PERSPECTIVES

6.1 CONCLUSION

In this dissertation, we discuss how to use the human gaze data to improve the performance of the weak supervised learning model in image classification. The background of this topic is in the era of rapidly growing information technology. As a consequence, the data to analyze is also growing dramatically. Since the amount of data that can be annotated by human can not keep up with the amount of data itself. the current well-developed supervised learning approach may confront bottlenecks in the future. In this context, the use of weak markings for high-performance learning methods is worthy of study.

Specifically, we try to solve the problem from two aspects: One is to propose a more time-saving annotation, human eye-tracking gaze, as an alternative annotation with respect to the traditional time-consuming annotation, *e.g.* bounding box. The other is to integrate gaze annotation into a weakly supervised learning scheme for image classification. This scheme benefits from the gaze annotation for inferring the regions containing the target object. An useful property of our model is that it only exploits gaze for training, while the test phase is gaze free. This property further reduces the demand of annotations. The two isolated aspects are connected together in our models, which further achieve competitive experimental results.

In addition to the innovation in terms of methodology, we also publicly release two datasets for promoting the research within the computer vision community. We first create a large scale food image dataset, UPMC Food-101. This large scale dataset is composed of about 100,000 recipes for a total of 101 food categories. Each item in this dataset is composed of an image and corresponding recipe text. As this dataset is multimodal, we try

to categorize the food with visual information, textual information, and `visual+textual` information. The fusion of multimodal data outperforms single information channel by a large margin. We then annotate part of the UPMC Food-101 with an eye-tracker to get a food-related gaze-based image dataset, UPMC-G20. We make various tests on this dataset to verify the generalization ability of our gaze-based weakly supervised models.

6.2 PERSPECTIVES

We develop the perspectives of this thesis from three aspects:

1. **Gaze analysis**. In Chapter 4 and 5, we incorporate gaze into the machine learning models as a weak supervision signal. In fact, there are many ways to exploit the gaze information in other forms of signal. A possible manner is to predict fixation from the image (Le Meur, Le Callet, and Barba 2007; Pan et al. 2016). The predicted fixations can be re-used for training the model, or for inferring the informative parts in the test images. Furthermore, attention-based deep learning models achieve success in various fields, including image captioning (Yang et al. 2016), visual question answering (Ben-Younes et al. 2017), machine translation (Vaswani et al. 2017). These researches show the powerful ability of processing the information in a selective manner. Up to now, most of the attention is inferred from the visual or textual data, only few researches is reported on using direct gaze-based attention.

2. **Weakly supervised learning**. In Chapter 4 and 5, our gaze-based WSL model achieves competitive result on the classification problem. We can naturally extend our model to other problems of computer vision, *e.g.* semantic segmentation, saliency, semantic boundaries, etc. This can be done by modifying the model to structural output (Yu and Joachims 2009; Durand, Thome, and Cord 2015). Also, we can achieve these goals by extending our model to an end-to-end deep learning model. By relating the various output to different loss function, a single deep learning model can do multi-tasks (Mordan et al. 2017; Durand et al. 2017; Durand, Thome, and Cord 2016) Besides, we can

inspire from the gaze attention mechanism to improve the modules, *e.g.* pooling layer (Sattar, Bulling, and Fritz 2016).

3. **Multimodal data processing**. In Chapter 3, we introduce a large scale food dataset UPMC Food-101. We can couple our gaze-based weakly classifiers with the deep-based textual processing strategies to build a more powerful multimodal classifier. Furthermore, we can study more methods for fusing the textual and visual data into the application of information retrieval, *e.g.* deep-based recipe retrieve system (Jingjing Chen 2016). Also, based on the relationship between the recipe data and the picture, we can study generating the food recipe from the food image (a.k.a image captioning (Vinyals et al. 2015)), or even generating the food image from the food recipe (Reed et al. 2016). The significance of this study is to let the machine really understand the relationship between different modal modes within multimodal data.

BIBLIOGRAPHY

AB, Tobii (2016). *Tobii Studio User's Manual Version 3.4.5* (cit. on p. 83).

Aizawa, K. and M. Ogawa (2015). "FoodLog: Multimedia Tool for Healthcare Applications." In: *IEEE MultiMedia* 22.2, pp. 4–8. DOI: 10.1109/MMUL.2015.39 (cit. on pp. 22, 23).

Aizawa, K., Y. Maruyama, H. Li, and C. Morikawa (2013). "Food Balance Estimation by Using Personal Dietary Tendencies in a Multimedia Food Log." In: *IEEE Transactions on Multimedia*. URL: http://dblp.uni-trier.de/db/journals/tmm/tmm15.html#AizawaMLM13 (cit. on pp. 22, 40).

Amano, Sosuke, Interdisciplinary Information, Kiyoharu Aizawa, and Makoto Ogawa (2014). "Frequency Statistics of Words Used in Japanese Food Records of FoodLog." In: *ACM UbiComp* (cit. on p. 23).

Amores, Jaume (2013). "Multiple Instance Classification: Review, Taxonomy and Comparative Study." In: *Artif. Intell.* 201, pp. 81–105. DOI: 10.1016/j.artint.2013.06.003. URL: http://dx.doi.org/10.1016/j.artint.2013.06.003 (cit. on p. 35).

Andrews, Stuart, Ioannis Tsochantaridis, and Thomas Hofmann (2002). "Support Vector Machines for Multiple-Instance Learning." In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 561–568 (cit. on p. 34).

Arandjelovic, Relja, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic (2016). "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307 (cit. on p. 6).

Avila, Sandra, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo Araujo (2012). "Pooling in Image Representation: the Visual Codeword Point of View." In: *Computer Vision and Image Understanding* (cit. on pp. 46, 54).

Azizpour, Hossein, Mostafa Arefiyan, Sobhan Naderi Parizi, and Stefan Carlsson (2015). "Spotlight the Negatives: A Generalized Discriminative Latent Model." In: *British Machine Vision Conference*, pp. 1–11 (cit. on pp. 35, 75, 76).

Babenko, Boris (2009). *Multiple Instance Learning: Algorithms and Applications*. Tech. rep. Dept. of Computer Science and Engineering University of California, San Diego (cit. on p. 35).

Bay, Herbert, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool (2008). "Speeded-Up Robust Features (SURF)." In: *Computer Vision and Image Understanding* 110.3. Similarity Matching in Computer Vision and Multimedia, pp. 346 – 359. DOI: http://doi.org/10.1016/j.cviu.2007.09.014. URL: http://www.sciencedirect.com/science/article/pii/S1077314207001555 (cit. on p. 15).

Beijbom, O., N. Joshi, D. Morris, S. Saponas, and S. Khullar (2015). "Menu-Match: Restaurant-Specific Food Logging from Images." In: *2015 IEEE Winter Conference on Applications of Computer Vision*, pp. 844–851. DOI: 10.1109/WACV.2015.117 (cit. on p. 22).

Ben-Younes, Hedi, Rémi Cadène, Nicolas Thome, and Matthieu Cord (2017). "MUTAN: Multimodal Tucker Fusion for Visual Question Answering." In: *arXiv preprint arXiv:1705.06676* (cit. on p. 96).

Bettadapura, Vinay, Edison Thomaz, Aman Parnami, Gregory D. Abowd, and Irfan Essa (2015). "Leveraging Context to Support Automated Food Recognition in Restaurants." In: *Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision*. WACV '15. Washington, DC, USA: IEEE Computer Society, pp. 580–587. DOI: 10.1109/WACV.2015.83. URL: http://dx.doi.org/10.1109/WACV.2015.83 (cit. on p. 22).

Bilen, Hakan, Vinay P. Namboodiri, and Luc J. Van Gool (2014). "Object and Action Classification with Latent Window Parameters." In: *Int. J. Comput. Vision* 106.3, pp. 237–251 (cit. on p. 35).

Bolaños, Marc, M Garolera, and P Radeva (2013). "Active labeling application applied to food-related object recognition." In: *ACM MM workshop* (cit. on p. 23).

Borji, Ali and Laurent Itti (2014). "Defending Yarbus: Eye movements reveal observers' task." In: *Journal of Vision* 14.3, p. 29. DOI: 10.1167/14.3.29. eprint: /data/journals/jov/932817/i1534-7362-14-3-29.pdf. URL: +http://dx.doi.org/10.1167/14.3.29 (cit. on p. 30).

Bossard, L. and et al. (2014). "Food-Mining – 101 Discriminative Components with Random Forests." In: *ECCV* (cit. on pp. 20, 21, 23, 39–41, 43, 48).

Bulling, Andreas, Jamie A. Ward, Hans Gellersen, and Gerhard Troster (2011). "Eye Movement Analysis for Activity Recognition Using Electrooculography." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 33.4, pp. 741–753. DOI: 10.1109/TPAMI.2010.86. URL: http://dx.doi.org/10.1109/TPAMI.2010.86 (cit. on p. 31).

Cadène, Rémi, Nicolas Thome, and Matthieu Cord (2016). "Master's Thesis : Deep Learning for Visual Recognition." In: *CoRR* abs/1610.05567. URL: http://arxiv.org/abs/1610.05567 (cit. on p. 47).

Carbonneau, Marc-André, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon (2016). "Multiple Instance Learning: A Survey of Problem Characteristics and Applications." In: *CoRR* abs/1612.03365. URL: http://arxiv.org/abs/1612.03365 (cit. on p. 35).

Chatfield, Ken, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman (2014). "Return of the Devil in the Details: Delving Deep into Convolutional Nets." In: *British Machine Vision Conference (BMVC)* (cit. on pp. 4, 19).

Chen, Jingjing, Lei Pang, and Chong-Wah Ngo (2017). "Cross-Modal Recipe Retrieval: How to Cook this Dish?" In: *MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part I*. Cham: Springer International Publishing, pp. 588–600. DOI: 10.1007/978-3-319-51811-4_48. URL: http://dx.doi.org/10.1007/978-3-319-51811-4_48 (cit. on p. 22).

Chen, M and et al. (2009). "PFID: Pittsburgh fast-food image dataset." In: *ICIP* (cit. on pp. 20, 21).

Christodoulidis, Stergios, Marios Anthimopoulos, and Stavroula G. Mougiakakou (2015). "Food Recognition for Dietary Assessment Using Deep Convolutional Neural Networks." In: *New Trends in Image Analysis and Processing - ICIAP 2015 Workshops - ICIAP 2015 International Workshops: BioFor, CTMR, RHEUMA, ISCA, MADiMa, SBMI, and QoEM, Genoa, Italy, September 7-8, 2015, Proceedings*, pp. 458–465. DOI: 10.1007/978-3-319-23222-5_56. URL: http://dx.doi.org/10.1007/978-3-319-23222-5_56 (cit. on p. 22).

Cisco, Systems Inc. (2016). *White paper: Cisco VNI Forecast and Methodology, 2015-2020*. Tech. rep. Cisco Systems Inc. (cit. on p. 1).

Cord, Matthieu and Philippe H Gosselin (2006). "Image retrieval using long-term semantic learning." In: *Image Processing, 2006 IEEE International Conference on*. IEEE, pp. 2909–2912 (cit. on p. 54).

Csurka, Gabriella, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray (2004). "Visual Categorization with Bags of Keypoints." In: *Workshop on Statistical Learning in Computer Vision. ECCV* (cit. on p. 15).

Dalal, N. and B. Triggs (2005). "Histograms of oriented gradients for human detection." In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177 (cit. on p. 15).

Damen, Dima, Teesid Leelasawassuk, and Walterio Mayol-Cuevas (2016). "You-Do, I-Learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance." In: *Computer Vision and Image Understanding* 149, pp. 98 –112 (cit. on p. 31).

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). "ImageNet: A Large-Scale Hierarchical Image Database." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255 (cit. on p. 4).

Deselaers, Thomas and Vittorio Ferrari (2010). "A Conditional Random Field for Multiple-Instance Learning." In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. Ed. by Johannes Fürnkranz and Thorsten Joachims. Omnipress, pp. 287–294. URL: http://www.icml2010.org/papers/87.pdf (cit. on p. 35).

Dietterich, Thomas G., Richard H. Lathrop, and Tomás Lozano-Pérez (1997). "Solving the multiple instance problem with axis-parallel rectangles." In: *Artificial Intelligence* 89, pp. 31–71 (cit. on p. 33).

Durand, Thibaut, Nicolas Thome, and Matthieu Cord (2015). "MANTRA: Minimum Maximum Latent Structural SVM for Image Classification and Ranking." In: *International Conference on Computer Vision*, pp. 2713–2721 (cit. on pp. 35, 75, 96).

Durand, Thibaut, Nicolas Thome, and Matthieu Cord (2016). "WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks." In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4743–4752 (cit. on pp. 35, 75, 96).

Durand, Thibaut, Nicolas Thome, Matthieu Cord, and David Picard (2014). "Incremental learning of latent structural SVM for weakly supervised image classification." In: *IEEE International Conference on Image Processing*, pp. 4246–4250 (cit. on p. 35).

Durand, Thibaut, Taylor Mordan, Nicolas Thome, and Matthieu Cord (2017). "WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on pp. 35, 75, 96).

Engelke, U., H. Liu, J. Wang, P. Le Callet, I. Heynderickx, H. J. Zepernick, and A. Maeder (2013). "Comparative Study of Fixation Density Maps." In: *IEEE Transactions on Image Processing* 22.3, pp. 1121–1133. DOI: 10.1109/TIP.2012.2227767 (cit. on p. 31).

Everingham, Mark, S. M. Ali Eslami, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman (2015). "The Pascal Visual Object Classes Challenge: A Retrospective." In: *International Journal of Computer Vision* 111.1, pp. 98–136 (cit. on pp. 5, 64, 67).

Fan, R. and et al. (2008). "LIBLINEAR: A library for large linear classification." In: *JMLR* (cit. on p. 45).

Farinella, G. M., M. Moltisanti, and S. Battiato (2014). "Classifying food images represented as Bag of Textons." In: *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 5212–5216. DOI: 10.1109/ICIP.2014.7026055 (cit. on p. 22).

Farinella, GM and et all (2014). "A Benchmark Dataset to Study Representation of Food Images." In: *ECCV workshop* (cit. on pp. 20, 21).

Fasel, Beat, Florent Monay, and Daniel Gatica-Perez (2004). "Latent Semantic Analysis of Facial Action Codes for Automatic Facial Expression Recognition." In: *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*. MIR '04. New York, NY, USA: ACM, pp. 181–188. DOI: 10.1145/1026711.1026742. URL: http://doi.acm.org/10.1145/1026711.1026742 (cit. on p. 15).

Fathi, Alireza, Yin Li, and James M. Rehg (2012). "Learning to Recognize Daily Actions Using Gaze." In: *European Conference on Computer Vision*, pp. 314–327 (cit. on pp. 31, 35, 56).

Fei-Fei, Li, Asha Iyer, Christof Koch, and Pietro Perona (2007). "What do we perceive in a glance of a real-world scene?" In: *Journal of Vision* 7, pp. 1–29 (cit. on p. 65).

Felzenszwalb, Pedro F., Ross B. Girshick, David A. McAllester, and Deva Ramanan (2010). "Object Detection with Discriminatively Trained Part-Based Models."

In: *IEEE Trans. Pattern Anal. Mach. Intell.* 32.9, pp. 1627–1645 (cit. on pp. 34, 56, 57).

Foulds, James and Eibe Frank (2010). "A review of multi-instance learning assumptions." In: *The Knowledge Engineering Review* 25.1, 1–25 (cit. on p. 35).

Fournier, Jérome, Matthieu Cord, and Sylvie Philipp-Foliguet (2001). "Back-propagation algorithm for relevance feedback in image retrieval." In: *Image Processing, 2001. Proceedings. 2001 International Conference on*. Vol. 1. IEEE, pp. 686–689 (cit. on p. 54).

Fukushima, Kunihiko (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position." In: *Biological Cybernetics* 36.4, pp. 193–202. DOI: 10.1007/BF00344251. URL: http://dx.doi.org/10.1007/BF00344251 (cit. on p. 4).

Garcez, A. S. d'Avila and G. Zaverucha (2012). "Multi-instance learning using recurrent neural networks." In: *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. DOI: 10.1109/IJCNN.2012.6252784 (cit. on p. 35).

Ge, Gary, Kiwon Yun, Dimitris Samaras, and Gregory J. Zelinsky (2015). "Action classification in still images using human eye movements." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 16–23 (cit. on pp. 31, 35, 56).

Gilani, Syed Omer, Ramanathan Subramanian, Yan Yan, David Melcher, Nicu Sebe, and Stefan Winkler (2015). "PET: An eye-tracking dataset for animal-centric Pascal object classes." In: *IEEE International Conference on Multimedia and Expo*, pp. 1–6 (cit. on p. 32).

Gkioxari, Georgia, Ross Girshick, and Jitendra Malik (2015). "Actions and Attributes from Wholes and Parts." In: *International Conference on Computer Vision (ICCV)*, pp. 2470–2478 (cit. on p. 86).

Gong, Yunchao, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik (2014). "Multi-scale Orderless Pooling of Deep Convolutional Activation Features." In: *European Conference on Computer Vision (ECCV)*, pp. 392–407 (cit. on p. 6).

Gordo, Albert, Adrien Gaidon, and Florent Perronnin (2015). "Deep Fishing: Gradient Features from Deep Nets." In: *British Machine Vision Conference*, pp. 1–12 (cit. on pp. 86, 87).

Gorisse, D., M. Cord, and F. Precioso (2011). "SALSAS: Sub-linear active learning strategy with approximate k-NN search." In: *Pattern Recognition* 44.10, pp. 2343–2357 (cit. on p. 54).

Gosselin, PH and M Cord (2008). "Active learning methods for interactive image retrieval." In: *Image Processing, IEEE Transactions on* 17.7, pp. 1200–1211 (cit. on p. 54).

Gosselin, Philippe Henri and Matthieu Cord (2004). "RETIN AL: An active learning strategy for image category retrieval." In: *Image Processing, 2004. ICIP'04. 2004 International Conference on*. Vol. 4. IEEE, pp. 2219–2222 (cit. on p. 54).

Gärtner, Thomas, Peter A. Flach, Adam Kowalczyk, and Alex J. Smola (2002). "Multi-Instance Kernels." In: *In Proc. 19th International Conf. on Machine Learning*. Morgan Kaufmann, pp. 179–186 (cit. on p. 33).

Hacisalihzade, Selim S., Lawrence W. Stark, and John S. Allen (1992). "Visual Perception and Sequences of Eye Movement Fixations: A Stochastic Modeling Approach." In: *IEEE Transactions on Systems, Man and Cybernetics* 22.3, pp. 474–481. DOI: 10.1109/21.155948 (cit. on p. 31).

Haji-Abolhassani, Amin and James J. Clark (2014). "An inverse Yarbus process: Predicting observers' task from eye movement patterns." In: *Vision Research* 103, pp. 127 –142. DOI: http://dx.doi.org/10.1016/j.visres.2014.08.014. URL: http://www.sciencedirect.com/science/article/pii/S0042698914002004 (cit. on p. 30).

Hassannejad, Hamid, Guido Matrella, Paolo Ciampolini, Ilaria De Munari, Monica Mordonini, and Stefano Cagnoni (2016). "Food Image Recognition Using Very Deep Convolutional Networks." In: *Proceedings of the 2Nd International Workshop on Multimedia Assisted Dietary Management*. MADiMa '16. Amsterdam, The Netherlands: ACM, pp. 41–49. DOI: 10.1145/2986035.2986042. URL: http://doi.acm.org/10.1145/2986035.2986042 (cit. on pp. 23, 40).

He, H., F. Kong, and J. Tan (2016). "DietCam: Multiview Food Recognition Using a Multikernel SVM." In: *IEEE Journal of Biomedical and Health Informatics* 20.3, pp. 848–855. DOI: 10.1109/JBHI.2015.2419251 (cit. on pp. 20–22, 40).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015a). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1026–1034 (cit. on p. 2).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015b). "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 37.9, pp. 1904–1916 (cit. on p. 6).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (cit. on pp. 4, 17, 18).

He, Y., C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp (2014). "Analysis of food images: Features and classification." In: *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 2744–2748. DOI: 10.1109/ICIP.2014.7025555 (cit. on pp. 20, 21).

Herranz, L., S. Jiang, and R. Xu (2017). "Modeling Restaurant Context for Food Recognition." In: *IEEE Transactions on Multimedia* 19.2, pp. 430–440. DOI: 10.1109/TMM.2016.2614861 (cit. on p. 22).

Herrera, Francisco, Sebastián Ventura, Rafael Bello, Chris Cornelis, Amelia Zafra, Dánel Sánchez-Tarragó, and Sarah Vluymans (2016). *Multiple instance learning : foundations and algorithms*. eng. Springer, pp. XI, 233. URL: http://dx.doi.org/10.1007/978-3-319-47759-6 (cit. on p. 35).

Hoai, Minh (2014). "Regularized max pooling for image categorization." In: *British Machine Vision Conference (BMVC)*, pp. 1–12 (cit. on p. 86).

Hoashi, H., T. Joutou, and K. Yanai (2010). "Image Recognition of 85 Food Categories by Feature Fusion." In: *2010 IEEE International Symposium on Multimedia*, pp. 296–301. DOI: 10.1109/ISM.2010.51 (cit. on p. 22).

Hubel, D. and T. N. Wiesel (1962). "Receptive Fields, Binocular Interaction, and Functional Architecture in the Cat's Visual Cortex." In: *Journal of Physiology* 160, pp. 106–154 (cit. on p. 4).

Huey, Edmund Burke (1908). *The psychology and pedagogy of reading*. Cambridge [Mass.] M.I.T. Press (cit. on p. 24).

Hussain, Zakria, Arto Klami, Jussi Kujala, Alex Po Leung, Kitsuchart Pasupa, Peter Auer, Samuel Kaski, Jorma Laaksonen, and John Shawe-Taylor (2014). "PinView: Implicit Feedback in Content-Based Image Retrieval." In: *CoRR* abs/1410.0471. URL: http://arxiv.org/abs/1410.0471 (cit. on p. 31).

Jack Hessel Nicolas Savva, Michael J. Wilber (2015). "Image Representations and New Domains in Neural Image Captioning." In: *EMNLP Vision + Learning workshop* (cit. on p. 22).

Jacob, Robert J. K. and Keith S. Karn (2003). "Eye Tracking in Human–Computer Interaction and Usability Research: Ready to Deliver the Promises." In: *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pp. 573–605 (cit. on p. 30).

Jacob, Robert J.K. (1995). "Eye Tracking in Advanced Interface Design." In: *Virtual environments and advanced interface design*. Oxford University Press, pp. 258–288 (cit. on p. 30).

Jingjing Chen, Chong-Wah Ngo (2016). "Deep-based Ingredient Recognition for Cooking Recipe Retrieval." In: *ACMMM* (cit. on pp. 9, 20–23, 40, 97).

Joachims, Thorsten, Thomas Finley, and Chun-Nam John Yu (2009). "Cutting-plane training of structural SVMs." In: *Machine Learning* 77.1, pp. 27–59 (cit. on p. 61).

Juneja, Mayank, Andrea Vedaldi, C. V. Jawahar, and Andrew Zisserman (2013). "Blocks That Shout: Distinctive Parts for Scene Classification." In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 923–930 (cit. on p. 34).

Just Marcel A. and Carpenter Patricia A. (1980). "A theory of reading: From eye fixations to comprehension." In: *Psychological Review* 87.4, pp. 329–354 (cit. on p. 24).

Kagaya, Hokuto, Kiyoharu Aizawa, and Makoto Ogawa (2014). "Food Detection and Recognition Using Convolutional Neural Network." In: *Proceedings of the 22Nd ACM International Conference on Multimedia*. MM '14. Orlando, Florida, USA: ACM, pp. 1085–1088. DOI: 10.1145/2647868.2654970. URL: http://doi.acm.org/10.1145/2647868.2654970 (cit. on p. 22).

Karthikeyan, S., V. Jagadeesh, R. Shenoy, M. Ecksteinz, and B. S. Manjunath (2013). "From Where and How to What We See." In: *International Conference on Computer Vision*, pp. 625–632 (cit. on pp. 31, 32, 35, 56).

Karthikeyan, S., Thuyen Ngo, Miguel P. Eckstein, and B. S. Manjunath (2015). "Eye tracking assisted extraction of attentionally important objects from videos." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3241–3250 (cit. on p. 31).

Kashlak, Adam B., Eoin Devane, Helge Dietert, and Henry Jackson (2017). "Markov models for ocular fixation locations in the presence and absence of colour." In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, n/a–n/a. DOI: 10.1111/rssc.12223. URL: http://dx.doi.org/10.1111/rssc.12223 (cit. on p. 31).

Kawano, Y. and K. Yanai (2014a). "Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation." In: *Proc. of ECCV Workshop on TASK-CV* (cit. on p. 41).

Kawano, Yoshiyuki and Keiji Yanai (2014b). "Food image recognition with deep convolutional features." In: *ACM UbiComp* (cit. on pp. 22, 23, 40).

Kawano, Yoshiyuki and Keiji Yanai (2014c). "FoodCam-256: A Large-scale Real-time Mobile Food RecognitionSystem employing High-Dimensional Features and Compression of Classifier Weights." In: *ACM International Conference on Multimedia*, pp. 761–762. DOI: 10.1145/2647868.2654869. URL: http://dl.acm.org/citation.cfm?doid=2647868.2654869 (cit. on pp. 20, 21, 23).

Kawano, Yoshiyuki and Keiji Yanai (2014d). "FoodCam-256: A Large-scale Real-time Mobile Food RecognitionSystem Employing High-Dimensional Features and Compression of Classifier Weights." In: *Proceedings of the 22Nd ACM International Conference on Multimedia*. MM '14. Orlando, Florida, USA: ACM, pp. 761–762. DOI: 10.1145/2647868.2654869. URL: http://doi.acm.org/10.1145/2647868.2654869 (cit. on pp. 20, 21).

Kawano, Yoshiyuki and Keiji Yanai (2015). "FoodCam: A real-time food recognition system on a smartphone." In: *Multimedia Tools and Applications* 74.14, pp. 5263–5287. DOI: 10.1007/s11042-014-2000-8. URL: http://dx.doi.org/10.1007/s11042-014-2000-8 (cit. on p. 23).

Keiji Yanai, Yoshiyuki Kawano (2015). "FOOD IMAGE RECOGNITION USING DEEP CONVOLUTIONAL NETWORK WITH PRE-TRAINING AND FINE-TUNING." In: *IEEE Internatinal Conference on Multimedia and Exposition, workshop CEA* (cit. on pp. 22, 23, 40).

Kesorn, K. and S. Poslad (2012). "An Enhanced Bag-of-Visual Word Vector Space Model to Represent Visual Content in Athletics Images." In: *IEEE Transactions on Multimedia* 14.1, pp. 211–222. DOI: 10.1109/TMM.2011.2170665 (cit. on p. 16).

Khanna, N. and et al. (2010). "An Overview of the Technology Assisted Dietary Assessment Project at Purdue University." In: *Proc. IEEE Int. Symp. Multimedia* (cit. on p. 24).

Kitamura, K., C. de Silva, T. Yamasaki, and K. Aizawa (2010). "Image processing based approach to food balance analysis for personal food logging." In: *2010 IEEE International Conference on Multimedia and Expo*, pp. 625–630. DOI: 10.1109/ICME.2010.5583021 (cit. on pp. 22, 40).

Klami, Arto, Craig Saunders, Teófilo Emídio de Campos, and Samuel Kaski (2008). "Can relevance of images be inferred from eye movements?" In: *Proceedings of*

the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2008, Vancouver, British Columbia, Canada, October 30-31, 2008 (cit. on p. 26).

Krafka, Kyle, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba (2016). "Eye Tracking for Everyone." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on p. 31).

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks." In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105 (cit. on pp. 4, 17, 18).

Kruthiventi, Srinivas S. S., Vennela Gudisa, Jaley H. Dholakiya, and R. Venkatesh Babu (2016). "Saliency Unified: A Deep Architecture for simultaneous Eye Fixation Prediction and Salient Object Segmentation." In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5781–5790 (cit. on p. 31).

Kumar, M. Pawan, Benjamin Packer, and Daphne Koller (2010). "Self-Paced Learning for Latent Variable Models." In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1189–1197 (cit. on p. 35).

Le, Quoc and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents." In: *ICML* (cit. on p. 51).

Le Callet, P. and E. Niebur (2013). "Visual Attention and Applications in Multimedia Technologies." In: *Proceedings of the IEEE* 101.9, pp. 2058–2067. DOI: 10.1109/JPROC.2013.2265801 (cit. on p. 30).

Le Meur, O. and P. Le Callet (2009). "What we see is most likely to be what matters: Visual attention and applications." In: *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 3085–3088. DOI: 10.1109/ICIP.2009.5414481 (cit. on p. 30).

Le Meur, O., P. Le Callet, D. Barba, and D. Thoreau (2006). "A coherent computational approach to model bottom-up visual attention." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.5, pp. 802–817. DOI: 10.1109/TPAMI.2006.86 (cit. on p. 30).

Le Meur, Olivier, Patrick Le Callet, and Dominique Barba (2007). "Predicting visual fixations on video based on low-level visual features." In: *Vision Research* 47.19, pp. 2483 –2498. DOI: http://dx.doi.org/10.1016/j.visres.2007.06.015. URL: http://www.sciencedirect.com/science/article/pii/S0042698907002593 (cit. on p. 96).

Learned-Miller, Erik, Gary B. Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua (2016). "Labeled Faces in the Wild: A Survey." In: *Advances in Face Detection and Facial Image Analysis*, pp. 189–248 (cit. on pp. 4, 17).

LeCun, Yann, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel (1989). "Backpropagation applied to handwritten zip code recognition." In: *Neural computation* 1.4, pp. 541–551 (cit. on pp. 4, 16, 18).

Li, Li jia, Hao Su, Li Fei-fei, and Eric P. Xing (2010). "Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification." In: *Advances in Neural Information Processing Systems*, pp. 1378–1386 (cit. on p. 66).

Li, Weixin and Nuno Vasconcelos (2015). "Multiple instance learning for soft bags via top instances." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4277–4285 (cit. on pp. 35, 73, 75, 78).

Li, Xiaolan and Afzal Godil (2010). "Investigating the Bag-of-words Method for 3D Shape Retrieval." In: *EURASIP J. Adv. Signal Process* 2010, 5:1–5:9. DOI: 10.1155/2010/108130. URL: http://dx.doi.org/10.1155/2010/108130 (cit. on p. 16).

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick (2014). "Microsoft COCO: Common Objects in Context." In: *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. Cham: Springer International Publishing, pp. 740–755. DOI: 10.1007/978-3-319-10602-1_48. URL: http://dx.doi.org/10.1007/978-3-319-10602-1_48 (cit. on p. 5).

Liu, Chang, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, and Yunsheng Ma (2016). "DeepFood: Deep Learning-Based Food Image Recognition for Computer-Aided Dietary Assessment." In: *Inclusive Smart Cities and Digital Health - 14th International Conference on Smart Homes and Health Telematics, ICOST 2016, Wuhan, China, May 25-27, 2016. Proceedings*, pp. 37–48. DOI: 10.1007/978-3-319-39601-9_4. URL: http://dx.doi.org/10.1007/978-3-319-39601-9_4 (cit. on pp. 23, 40).

Lopez, Stephanie, Arnaud Revel, Diane Lingrand, and Frédéric Precioso (2015). "One gaze is worth ten thousand (key-)words." In: *IEEE International Conference on Image Processing (ICIP)*, pp. 3150–3154 (cit. on pp. 31, 32).

Lowe, David G. (2004). "Distinctive Image Features from Scale-Invariant Keypoints." In: *Int. J. Comput. Vision* 60.2, pp. 91–110 (cit. on pp. 3, 15).

Luis Herranz Ruihan Xu, Shuqiang Jiang (2015). "A PROBABILISTIC MODEL FOR FOOD IMAGE RECOGNITION IN RESTAURANTS." In: *IEEE Internatinal Conference on Multimedia and Exposition* (cit. on pp. 20–22).

L.Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus (2013). "Regularization of Neural Networks using DropConnect." In: *ICML* (cit. on p. 19).

Ma, Wei-Ying and B. S. Manjunath (1999). "NeTra: A Toolbox for Navigating Large Image Databases." In: *Multimedia Syst.* 7.3, pp. 184–198. DOI: 10.1007/s005300050121. URL: http://dx.doi.org/10.1007/s005300050121 (cit. on p. 3).

Majaranta, Päivi and Andreas Bulling (2014). "Eye Tracking and Eye-Based Human–Computer Interaction." In: *Advances in Physiological Computing*. Ed. by Stephen H. Fairclough and Kiel Gilleade. London: Springer London, pp. 39–65. DOI: 10.1007/978-1-4471-6392-3_3. URL: http://dx.doi.org/10.1007/978-1-4471-6392-3_3 (cit. on p. 30).

Maron, O. and T. Lozano-Perez (1998a). "A framework for multiple-instance learning." In: *Advances in Neural Information Processing Systems (NIPS)* (cit. on p. 33).

Maron, O. and T. Lozano-Perez (1998b). "Multiple-Instance Learning for Natural Scene Classification." In: *ICML* (cit. on p. 33).

Mathe, Stefan, Aleksis Pirinen, and Cristian Sminchisescu (2016). "Reinforcement Learning for Visual Object Detection." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2894–2902 (cit. on pp. 35, 56).

Mathe, Stefan and Cristian Sminchisescu (2013). "Action from Still Image Dataset and Inverse Optimal Control to Learn Task Specific Visual Scanpaths." In: *Advances in Neural Information Processing Systems*, pp. 1923–1931 (cit. on pp. 26, 31, 32, 63).

Mathe, Stefan and Cristian Sminchisescu (2014). "Multiple Instance Reinforcement Learning for Efficient Weakly-Supervised Detection in Images." In: *CoRR* abs/1412.0100 (cit. on pp. 35, 56).

Mathe, Stefan and Cristian Sminchisescu (2015). "Actions in the Eye: Dynamic Gaze Datasets and Learnt Saliency Models for Visual Recognition." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 37.7, pp. 1408–1424 (cit. on p. 31).

Matsuda, Y and K Yanai (2012). "Multiple-food recognition considering co-occurrence employing manifold ranking." In: *ICPR* (cit. on p. 22).

Matsunaga, Hiroki, Keisuke Doman, Takatsugu Hirayama, Ichiro Ide, Daisuke Deguchi, and Hiroshi Murase (2015). "Tastes and Textures Estimation of Foods Based on the Analysis of Its Ingredients List and Image." In: *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops: ICIAP 2015 International Workshops, BioFor, CTMR, RHEUMA, ISCA, MADiMa, SBMI, and QoEM, Genoa, Italy, September 7-8, 2015, Proceedings*. Cham: Springer International Publishing, pp. 326–333. DOI: 10.1007/978-3-319-23222-5_40. URL: http://dx.doi.org/10.1007/978-3-319-23222-5_40 (cit. on p. 22).

Matthew Blaschko, Pawan Kumar, Ben Taskar (2013). "Tutorial: Visual Learning with Weak Supervision." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on pp. 7, 33).

Michael Posner (1980). "ORIENTING OF ATTENTION." In: *The Quarterly Journal of Experimental Psychology* 32, pp. 3–25 (cit. on p. 24).

Mikolov, T. and et all (2013). "Distributed representations of words and phrases and their compositionality." In: *NIPS* (cit. on pp. 40, 50).

Min, Weiqing, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz (2016). "Being a Super Cook: Joint Food Attributes and Multi-Modal Content Modeling for Recipe Retrieval and Exploration." In: *IEEE Transactions on Multimedia* X.XX, pp. 1–1. DOI: 10.1109/TMM.2016.2639382. URL: http://ieeexplore.ieee.org/document/7782829/ (cit. on pp. 9, 20–22).

Mishra, Ajay K., Yiannis Aloimonos, and Loong Fah Cheong (2009). "Active segmentation with fixation." In: *IEEE International Conference on Computer Vision*, pp. 468–475 (cit. on p. 31).

Miyazaki, T., G. C. de Silva, and K. Aizawa (2011). "Image-based Calorie Content Estimation for Dietary Assessment." In: *2011 IEEE International Symposium on Multimedia*, pp. 363–368. DOI: 10.1109/ISM.2011.66 (cit. on p. 22).

Mordan, Taylor, Nicolas Thome, Gilles Henaff, and Matthieu Cord (2017). "Deformable Part-based Fully Convolutional Network for Object Detection." In: *Proceedings of the British Machine Vision Conference (BMVC)* (cit. on p. 96).

Myers, Austin, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin Murphy (2016). "Im2Calories: Towards an automated mobile vision food diary." In: *Proceedings of the IEEE International Conference on Computer*

Vision. Vol. 11-18-Dece, pp. 1233–1241. DOI: 10.1109/ICCV.2015.146. URL: https://www.cs.ubc.ca/{~}murphyk/Papers/im2calories{_}iccv15.pdf (cit. on pp. 20–23, 40).

Ninassi, A., O. Le Meur, P. Le Callet, and D. Barba (2007). "Does where you Gaze on an Image Affect your Perception of Quality? Applying Visual Attention to Image Quality Metric." In: *2007 IEEE International Conference on Image Processing*. Vol. 2, pp. II –169–II –172. DOI: 10.1109/ICIP.2007.4379119 (cit. on p. 30).

Ninassi, A., O. Le Meur, P. Le Callet, and D. Barba (2009). "Considering Temporal Variations of Spatial Visual Distortions in Video Quality Assessment." In: *IEEE Journal of Selected Topics in Signal Processing* 3.2, pp. 253–265. DOI: 10.1109/JSTSP.2009.2014806 (cit. on p. 30).

Noronha, Jon, Eric Hysen, Haoqi Zhang, and Krzysztof Z. Gajos (2011). "Platemate: Crowdsourcing Nutritional Analysis from Food Photographs." In: *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. UIST '11. Santa Barbara, California, USA: ACM, pp. 1–12. DOI: 10.1145/2047196.2047198. URL: http://doi.acm.org/10.1145/2047196.2047198 (cit. on p. 23).

Oliveira, Luciano, Victor Costa, Gustavo Neves, Talmai Oliveira, Eduardo Jorge, and Miguel Lizarraga (2014). "A mobile, lightweight, poll-based food identification system." In: *Pattern Recognition* 47.5, pp. 1941 –1952. DOI: http://dx.doi.org/10.1016/j.patcog.2013.12.006. URL: http://www.sciencedirect.com/science/article/pii/S0031320313005244 (cit. on pp. 22, 23).

Olsen, Anneli (2012). *The Tobii I-VT Fixation Filter* (cit. on p. 80).

Oquab, Maxime, Léon Bottou, Ivan Laptev, and Josef Sivic (2014). "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks." In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1717–1724 (cit. on pp. 6, 86, 87).

Pan, Junting, Elisa Sayrol, Xavier Giró i Nieto, Kevin McGuinness, and Noel E. O'Connor (2016). "Shallow and Deep Convolutional Networks for Saliency Prediction." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 598–606 (cit. on pp. 31, 96).

Pandey, Megha and Svetlana Lazebnik (2011). "Scene recognition and weakly supervised object localization with deformable part-based models." In: *IEEE International Conference on Computer Vision*, pp. 1307–1314 (cit. on p. 34).

Papadopoulos, Dim P., Alasdair D. F. Clarke, Frank Keller, and Vittorio Ferrari (2014). "Training Object Class Detectors from Eye Tracking Data." In: *European Conference on Com- puter Vision (ECCV)*, pp. 361–376 (cit. on pp. 7, 31, 32, 35, 56, 63, 83).

Papadopoulos, G. T., K. C. Apostolakis, and P. Daras (2014). "Gaze-Based Relevance Feedback for Realizing Region-Based Image Retrieval." In: *IEEE Transactions on Multimedia* 16.2, pp. 440–454. DOI: 10.1109/TMM.2013.2291535 (cit. on p. 31).

Picard, D., M. Cord, and A. Revel (2008). "Image retrieval over networks: Active learning using ant algorithm." In: *Multimedia, IEEE Transactions on* 10.7, pp. 1356–1365 (cit. on p. 54).

Pouladzadeh, P., S. Shirmohammadi, and R. Al-Maghrabi (2014). "Measuring Calorie and Nutrition From Food Image." In: *IEEE Transactions on Instrumentation and Measurement* 63.8, pp. 1947–1956. DOI: 10.1109/TIM.2014.2303533 (cit. on p. 22).

Pouladzadeh, Parisa, Abdulsalam Yassine, and Shervin Shirmohammadi (2015). "FooDD: Food Detection Dataset for Calorie Measurement Using Food Images." In: *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops: ICIAP 2015 International Workshops, BioFor, CTMR, RHEUMA, ISCA, MADiMa, SBMI, and QoEM, Genoa, Italy, September 7-8, 2015, Proceedings*. Cham: Springer International Publishing, pp. 441–448. DOI: 10.1007/978-3-319-23222-5_54. URL: http://dx.doi.org/10.1007/978-3-319-23222-5_54 (cit. on pp. 20–22).

Pushmeet Kohli and L'ubor Ladický and Philip H.S. Torr (2009). "Robust Higher Order Potentials for Enforcing Label Consistency." In: *Int. J. Comput. Vision* 82.3, pp. 302–324 (cit. on p. 31).

Ramanathan, Subramanian, Victoria Yanulevskaya, and Nicu Sebe (2011). "Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements." In: *International Conference on Multimedia*, pp. 33–42 (cit. on p. 30).

Ramanathan, Subramanian, Harish Katti, Nicu Sebe, Mohan S. Kankanhalli, and Tat-Seng Chua (2010). "An Eye Fixation Database for Saliency Detection in Images." In: *European Conference on Computer Vision*, pp. 30–43 (cit. on p. 31).

Reed, Scott, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee (2016). "Generative Adversarial Text-to-Image Synthesis." In: *Proceedings of The 33rd International Conference on Machine Learning* (cit. on p. 97).

Ren, Weiqiang, Kaiqi Huang, Dacheng Tao, and Tieniu Tan (2016). "Weakly Supervised Large Scale Object Localization with Multiple Instance Learning and Bag Splitting." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.2, pp. 405–416. DOI: 10.1109/TPAMI.2015.2456908 (cit. on pp. 34, 35).

Rohrbach, Marcus and S Amin (2012). "A database for fine grained activity detection of cooking activities." In: *CVPR* (cit. on p. 21).

Rosenblatt, Frank (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory (cit. on pp. 4, 16).

Russakovsky, Olga, Yuanqing Lin, Kai Yu, and Li Fei-Fei (2012a). "Object-Centric Spatial Pooling for Image Classification." In: *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–15. DOI: 10.1007/978-3-642-33709-3_1. URL: http://dx.doi.org/10.1007/978-3-642-33709-3_1 (cit. on p. 5).

Russakovsky, Olga, Yuanqing Lin, Kai Yu, and Fei-Fei Li (2012b). "Object-Centric Spatial Pooling for Image Classification." In: *European Conference on Computer Vision*, pp. 1–15 (cit. on p. 35).

Russakovsky, Olga, Amy L. Bearman, Vittorio Ferrari, and Fei-Fei Li (2016). "What's the point: Semantic segmentation with point supervision." In: *ECCV* (cit. on p. 7).

Salton, Gerard and Michael J. McGill (1986). *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc. (cit. on p. 3).

Salvador, Amaia, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba (2017). "Learning Cross-modal Embeddings for Cooking Recipes and Food Images." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (cit. on p. 22).

Saon, G., T. Sercu, S. J. Rennie, and H. J. Kuo (2016). "The IBM 2016 English conversational telephone speech recognition system." In: *Interspeech*, pp. 7–11 (cit. on pp. 4, 17).

Sattar, Hosnieh, Andreas Bulling, and Mario Fritz (2016). "Predicting the Category and Attributes of Mental Pictures Using Deep Gaze Pooling." In: *CoRR* abs/1611.10162. URL: http://arxiv.org/abs/1611.10162 (cit. on pp. 31, 97).

Sattar, Hosnieh, Sabine Muller, Mario Fritz, and Andreas Bulling (2015). "Prediction of Search Targets From Fixations in Open-World Settings." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on pp. 26, 30).

Schroff, F., A. Criminisi, and A. Zisserman (2011). "Harvesting Image Databases from the Web." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.4, pp. 754–766. DOI: 10.1109/TPAMI.2010.133 (cit. on p. 41).

Shapovalova, Nataliya, Michalis Raptis, Leonid Sigal, and Greg Mori (2013). "Action is in the Eye of the Beholder: Eye-gaze Driven Model for Spatio-Temporal Action Localization." In: *Advances in Neural Information Processing Systems*, pp. 2409–2417 (cit. on pp. 31, 35, 75).

Shcherbatyi, Iaroslav, Andreas Bulling, and Mario Fritz (2015). "GazeDPM: Early Integration of Gaze Information in Deformable Part Models." In: *CoRR* abs/1505.05753 (cit. on pp. 31, 35, 56).

Shen, Wei, Xiang Bai, Zihao Hu, and Zhijiang Zhang (2016). "Multiple instance subspace learning via partial random projection tree for local reflection symmetry in natural images." In: *Pattern Recognition* 52, pp. 306–316 (cit. on p. 34).

Shrivastava, Ashish, Vishal M. Patel, Jaishanker K. Pillai, and Rama Chellappa (2015). "Generalized Dictionaries for Multiple Instance Learning." In: *Int. J. Comput. Vision* 114.2-3, pp. 288–305 (cit. on p. 34).

Silver David, Huang Aja, and et.al (2016). "Mastering the game of Go with deep neural networks and tree search." In: *Nature* 529.7587, pp. 484–489 (cit. on pp. 4, 17).

Simonyan, K and A Zisserman (2015). "Very deep convolutional networks for large-scale image recognition." In: *ICLR* (cit. on pp. 18, 40, 47).

Sivic, J. and A. Zisserman (2003). "Video Google: A Text Retrieval Approach to Object Matching in Videos." In: *International Conference on Computer Vision (ICCV)* (cit. on pp. 3, 15).

Smeulders, Arnold W. M., Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain (2000). "Content-Based Image Retrieval at the End of the Early Years." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 22.12, pp. 1349–1380. DOI: 10.1109/34.895972. URL: http://dx.doi.org/10.1109/34.895972 (cit. on p. 2).

Song, Zheng, Qiang Chen, ZhongYang Huang, Yang Hua, and Shuicheng Yan (2011). "Contextualizing object detection and classification." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1585–1592 (cit. on pp. 86, 87).

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." In: *J. Mach. Learn. Res* 15, pp. 1929–1958 (cit. on p. 19).

Steil, Julian and Andreas Bulling (2015). "Discovery of Everyday Human Activities from Long-term Visual Behaviour Using Topic Models." In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp '15. Osaka, Japan: ACM, pp. 75–85. DOI: 10.1145/2750858.2807520. URL: http://doi.acm.org/10.1145/2750858.2807520 (cit. on p. 31).

Stein, S. and McKenna, S. J. (2013). "User-adaptive Models for Recognizing Food Preparation Activities." In: *ACM MM workshop CEA* (cit. on p. 21).

Su, Han, Ting-Wei Lin, Cheng-Te Li, Man-Kwan Shan, and Janet Chang (2014). "Automatic Recipe Cuisine Classification by Ingredients." In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. UbiComp '14 Adjunct. Seattle, Washington: ACM, pp. 565–570. DOI: 10.1145/2638728.2641335. URL: http://doi.acm.org/10.1145/2638728.2641335 (cit. on p. 22).

Su, Hao, Jia Deng, and Li Fei-Fei (2012). "Crowdsourcing Annotations for Visual Object Detection." In: *AAAI Workshop*, pp. 1–6 (cit. on p. 31).

Sun, Jian and Jean Ponce (2013). "Learning Discriminative Part Detectors for Image Classification and Cosegmentation." In: *International Conference on Computer Vision (ICCV)*, pp. 3400–3407 (cit. on p. 34).

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich (2015). "Going Deeper with Convolutions." In: *Computer Vision and Pattern Recognition (CVPR)* (cit. on pp. 18, 40).

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna (2016). "Rethinking the Inception Architecture for Computer Vision." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308. URL: http://dx.doi.org/10.1109/CVPR.2016.308 (cit. on p. 40).

Toldo, Roberto, Umberto Castellani, and Andrea Fusiello (2009). "A Bag of Words Approach for 3D Object Categorization." In: *Proceedings of the 4th International Conference on Computer Vision/Computer Graphics CollaborationTechniques*. MIRAGE '09. Rocquencourt, France: Springer-Verlag, pp. 116–127. DOI: 10.1007/978-3-642-01811-4_11. URL: http://dx.doi.org/10.1007/978-3-642-01811-4_11 (cit. on p. 16).

Vapnik, Vladimir and Rauf Izmailov (2015). "Learning Using Privileged Information: Similarity Control and Knowledge Transfer." In: *J. Mach. Learn. Res* 16, pp. 2023–2049 (cit. on p. 63).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention Is All You Need." In: *Arxiv* (cit. on p. 96).

Vig, Eleonora, Michael Dorr, and David D. Cox (2012). "Saliency-based selection of sparse descriptors for action recognition." In: *IEEE International Conference on Image Processing (ICIP)* (cit. on p. 26).

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan (2015). "Show and Tell: A Neural Image Caption Generator." In: *CVPR* (cit. on p. 97).

W. Susanto, M. Rohrbach and B. Schiele (2012). "3d object detection with multiple kinects." In: *Computer Vision - ECCV 2012. Workshops and Demonstrations* (cit. on p. 21).

Walber, Tina, Ansgar Scherp, and Steffen Staab (2013). "Can You See It? Two Novel Eye-Tracking-Based Measures for Assigning Tags to Image Regions." In: *Advances in Multimedia Modeling, International Conference*, pp. 36–46 (cit. on p. 31).

Wang, Hsueh-Cheng and Marc Pomplun (2012). "The attraction of visual attention to texts in real-world scenes." In: *Journal of Vision* 12, pp. 1–17 (cit. on p. 80).

Wang, J., M. P. Da Silva, P. Le Callet, and V. Ricordel (2013a). "Computational Model of Stereoscopic 3D Visual Saliency." In: *IEEE Transactions on Image Processing* 22.6, pp. 2151–2165. DOI: 10.1109/TIP.2013.2246176 (cit. on p. 31).

Wang, Jingyan, Yongping Li, Ying Zhang, Chao Wang, Honglan Xie, Guoling Chen, and Xin Gao (2011). "Bag-of-Features Based Medical Image Retrieval via Multiple Assignment and Visual Words Weighting." In: *IEEE Trans. Med. Imaging* 30.11, pp. 1996–2011. DOI: 10.1109/TMI.2011.2161673. URL: http://dx.doi.org/10.1109/TMI.2011.2161673 (cit. on p. 15).

Wang, Jun and Jean-Daniel Zucker (2000). "Solving the Multiple-Instance Problem: A Lazy Learning Approach." In: *Proceedings of the Seventeenth International Conference on Machine Learning*. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 1119–1126. URL: http://dl.acm.org/citation.cfm?id=645529.757771 (cit. on p. 33).

Wang, Junle, Damon M. Chandler, and Patrick Le Callet (2010). "Quantifying the relationship between visual salience and visual importance." In: vol. 7527, 75270K–75270K–9. DOI: 10.1117/12.845231. URL: http://dx.doi.org/10.1117/12.845231 (cit. on p. 30).

Wang, Xin, Nicolas Thome, and Matthieu Cord (2016). "Gaze latent support vector machine for image classification." In: *IEEE International Conference on Image Processing (ICIP)*, pp. 236–240 (cit. on pp. iii, 12, 55, 86, 87).

Wang, Xin, Nicolas Thome, and Matthieu Cord (2017). "Gaze Latent Support Vector Machine for Image Classification Improved by Weakly Supervised Region Selection." In: *Pattern Recognition*, pp. –. DOI: https://doi.org/10.1016/j.patcog.2017.07.001. URL: http://www.sciencedirect.com/science/article/pii/S0031320317302625 (cit. on pp. iii, 12, 74).

Wang, Xin, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso (2015a). "Recipe recognition with large multimodal food dataset." In: *IEEE International Conference on Multimedia & Expo Workshops*, pp. 1–6 (cit. on pp. 12, 21, 32, 80).

Wang, Xinggang, Baoyuan Wang, Xiang Bai, Wenyu Liu, and Zhuowen Tu (2013b). "Max-Margin Multiple-Instance Dictionary Learning." In: *International Conference on Machine Learning*, pp. 846–854 (cit. on p. 34).

Wang, Xinggang, Zhuotun Zhu, Cong Yao, and Xiang Bai (2015b). "Relaxed Multiple-Instance SVM with Application to Object Discovery." In: *International Conference on Computer Vision (ICCV)*, pp. 1224–1232 (cit. on p. 34).

Wang, Xinggang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu (2016). "Revisiting Multiple Instance Neural Networks." In: *arxiv*, pp. 1–9 (cit. on p. 33).

Wang, Y. and G. Mori (2009). "Human Action Recognition by Semilatent Topic Models." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.10, pp. 1762–1774. DOI: 10.1109/TPAMI.2009.43 (cit. on p. 15).

Wu, Wen and Jie Yang (2009). "Fast food recognition from videos of eating for calorie estimation." In: *2009 IEEE International Conference on Multimedia and Expo*, pp. 1210–1213. DOI: 10.1109/ICME.2009.5202718 (cit. on p. 22).

Xie, H., L. Yu, and Q. Li (2010). "A Hybrid Semantic Item Model for Recipe Search by Example." In: *2010 IEEE International Symposium on Multimedia*, pp. 254–259. DOI: 10.1109/ISM.2010.44 (cit. on p. 22).

Xu, Jia, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M. Rehg, and Vikas Singh (2015a). "Gaze-enabled egocentric video summarization via constrained submodular maximization." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2235–2244 (cit. on p. 31).

Xu, R., L. Herranz, S. Jiang, S. Wang, X. Song, and R. Jain (2015b). "Geolocalized Modeling for Dish Recognition." In: *IEEE Transactions on Multimedia* 17.8, pp. 1187–1199. DOI: 10.1109/TMM.2015.2438717 (cit. on p. 22).

Xu, Ruihan, Luis Herranz, Shuqiang Jiang, Shuang Wang, Xinhang Song, and Ramesh Jain (2015c). "Geolocalized Modeling for Dish Recognition." In: *IEEE Transactions on Multimedia* 17.8, pp. 1187–1199. DOI: 10.1109/TMM.2015.2438717 (cit. on pp. 20, 21).

Yang, S., M. Chen, D. Pomerleau, and R. Sukthankar (2010). "Food recognition using statistics of pairwise local features." In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2249–2256. DOI: 10.1109/CVPR.2010.5539907 (cit. on pp. 22, 40).

Yang, Zhilin, Ye Yuan, Yuexin Wu, Ruslan Salakhutdinov, and William W. Cohen (2016). "Review Networks for Caption Generation." In: *NIPS* (cit. on p. 96).

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York: Plenum., p. 190 (cit. on pp. 24, 25).

Ye, P. and D. Doermann (2012). "No-Reference Image Quality Assessment Using Visual Codebooks." In: *IEEE Transactions on Image Processing* 21.7, pp. 3129–3138. DOI: 10.1109/TIP.2012.2190086 (cit. on p. 16).

Yosinski, Jason, Jeff Clune, Yoshua Bengio, and Hod Lipson (2014). "How transferable are features in deep neural networks?" In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 3320–3328 (cit. on pp. 4, 19).

Yu, Chun-Nam John and Thorsten Joachims (2009). "Learning structural SVMs with latent variables." In: *ICML* (cit. on p. 96).

Yuille, Alan L. and Anand Rangarajan (2001). "The Concave-Convex Procedure (CCCP)." In: *Advances in Neural Information Processing Systems 14*, pp. 1033–1040 (cit. on p. 62).

Yun, Kiwon, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg (2013). "Studying Relationships between Human Gaze, Description, and Computer Vision." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 739–746 (cit. on pp. 7, 26, 30, 31).

Zeiler, Matthew D. and Rob Fergus (2014). "Visualizing and Understanding Convolutional Networks." In: *European Conference on Computer Vision*, pp. 818–833 (cit. on pp. 86, 87).

Zelinsky, Gregory J., Yifan Peng, and Dimitris Samaras (2013). "Eye can read your mind: Decoding gaze fixations to reveal categorical search targets." In: *Journal of Vision* 13.14, p. 10. DOI: 10.1167/13.14.10. eprint: /data/journals/jov/933540/i1534-7362-13-14-10.pdf. URL: +http://dx.doi.org/10.1167/13.14.10 (cit. on p. 31).

Zhang, Min-Ling and Zhi-Hua Zhou (2004). "Improve Multi-Instance Neural Networks through Feature Selection." In: *Neural Processing Letters* 19.1, pp. 1–10. DOI: 10.1023/B:NEPL.0000016836.03614.9f. URL: http://dx.doi.org/10.1023/B:NEPL.0000016836.03614.9f (cit. on p. 33).

Zhang, Qi and Sally A. Goldman (2001). "EM-DD: An Improved Multiple-Instance Learning Technique." In: *In Advances in Neural Information Processing Systems*. MIT Press, pp. 1073–1080 (cit. on p. 33).

Zhang, W., A. Borji, Z. Wang, P. Le Callet, and H. Liu (2016). "The Application of Visual Saliency Models in Objective Image Quality Assessment: A Statistical Evaluation." In: *IEEE Transactions on Neural Networks and Learning Systems* 27.6, pp. 1266–1278. DOI: 10.1109/TNNLS.2015.2461603 (cit. on p. 30).

Zhou, Feng and Yuanqing Lin (2016). "Fine-Grained Image Classification by Exploring Bipartite-Graph Labels." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 1124–1133. DOI: 10.1109/CVPR.2016.127. URL: http://dx.doi.org/10.1109/CVPR.2016.127 (cit. on p. 22).

Zhou, Jie, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu (2016). "Deep Recurrent Models with Fast-Forward Connections for Neural Machine Translation." In:

Transactions of the Association for Computational Linguistics (TACL) 4, pp. 371–383 (cit. on pp. 4, 17).

Zhou, Zhi-Hua (2004). *Multi-Instance Learning: A Survey*. Tech. rep. National Laboratory for Novel Software Technology (cit. on p. 35).

Zhou, Zhi-Hua, Yu-Yin Sun, and Yu-Feng Li (2009). "Multi-instance Learning by Treating Instances As non-I.I.D. Samples." In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. Montreal, Quebec, Canada: ACM, pp. 1249–1256. DOI: 10.1145/1553374.1553534. URL: http://doi.acm.org/10.1145/1553374.1553534 (cit. on p. 35).

Zhou, Zhi-Hua and Min-Ling Zhang (2003). "Ensembles of Multi-instance Learners." In: *Machine Learning: ECML 2003: 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 492–502. DOI: 10.1007/978-3-540-39857-8_44. URL: http://dx.doi.org/10.1007/978-3-540-39857-8_44 (cit. on p. 33).