Machine learning solutions to visual recognition problems

Jakob Verbeek

HAL Id: tel-01343391

https://inria.hal.science/tel-01343391

Submitted on 8 Jul 2016

Machine learning solutions
to visual recognition problems

Jakob Verbeek

Summary

This thesis gives an overview of my research since my arrival in December 2005 as a postdoctoral fellow at the in the LEAR team at INRIA Rhône-Alpes. After a general introduction in Chapter 1, the contributions are presented in chapters 2–4 along three themes. In each chapter we describe the contributions, their relation to related work, and highlight two contributions with more detail.

Chapter 2 is concerned with contributions related to the Fisher vector representation. We highlight an extension of the representation based on modeling dependencies among local descriptors (Cinbis et al., 2012, 2016a). The second highlight is on an approximate normalization scheme which speeds-up applications for object and action localization (Oneata et al., 2014b).

In Chapter 3 we consider the contributions related to metric learning. The first contribution we highlight is a nearest-neighbor based image annotation method that learns weights over neighbors, and effectively determines the number of neighbors to use (Guillaumin et al., 2009a). The second contribution we highlight is an image classification method based on metric learning for the nearest class mean classifier that can efficiently generalize to new classes (Mensink et al., 2012, 2013b).

The third set of contributions, presented in Chapter 4, is related to learning visual recognition models from incomplete supervision. The first highlighted contribution is an interactive image annotation method that exploits dependencies across different image labels, to improve predictions and to identify the most informative user input (Mensink et al., 2011, 2013a). The second highlighted contribution is a multi-fold multiple instance learning method for learning object localization models from training images where we only know if the object is present in the image or not (Cinbis et al., 2014, 2016b).

Finally, Chapter 5 summarizes the contributions, and presents future research directions. A curriculum vitae with a list of publications is available in Appendix A.

Résumé

Cette thèse donne un aperçu de mes recherches depuis mon arrivée en décembre 2005 en tant que postdoctorat au sein de l'équipe LEAR à l'INRIA Rhône-Alpes. Après une introduction générale au Chapitre 1, les contributions seront présentées dans les chapitres 2–4. Chaque chapitre décrira les contributions liés à un thème et leur relation avec les travaux y afférent. Deux contributions seront également mise en exergue.

Le Chapitre 2 concernera les contributions liées à la représentation vectorielle de Fisher. Nous mettons en avant une extension de cette représentation basée sur la modélisation des dépendances parmi les descripteurs locaux (Cinbis et al., 2012, 2016a). La deuxième contribution présentée en détail est un ensemble d'approximations des normalisations du vecteur de Fisher, qui permettent une accélération dans des applications de localisation d'objets et d'actions (Oneata et al., 2014b).

Dans le Chapitre 3, nous considérerons les contributions liées à l'apprentissage de métrique. La première contribution que nous détaillerons est une méthode d'annotation d'image type plus proche voisin. Cette méthode permet d'affecter des poids aux voisins et de déterminer le nombre de voisins à utiliser (Guillaumin et al., 2009a). La deuxième contribution que nous mettrons en valeur est une méthode de classification d'image basée sur l'apprentissage de métrique qui permet de généraliser à de nouvelles classes (Mensink et al., 2012, 2013b).

La troisième série de contributions, présentées dans le Chapitre 4, sont liées à l'apprentissage de modèles de reconnaissance visuelle avec des données incomplètes. La contribution mise en valeur est une méthode d'annotation d'image interactive qui exploite les dépendances entre les différentes étiquettes d'image, pour améliorer les prévisions et optimiser les interactions avec l'utilisateur (Mensink et al., 2011, 2013a). La deuxième contribution majeure est une méthode d'appentissage à multiple-instances pour apprendre des modèles de localisation d'objet à partir d'images pour lesquelles nous savons seulement si l'objet est présent dans l'image ou non (Cinbis et al., 2014, 2016b).

Enfin, le Chapitre 5 résume les contributions et présente des pistes pour de futures recherches. Une curriculum vitae avec une liste des publications est disponible en Annexe A.

Contents

Chapter 1

Introduction

In this chapter we briefly sketch the context of the work presented in this document in Section 1.1. Then, in Section 1.2 and briefly describe the content of the rest of the document.

1.1 Context

In the last decade we have witnessed an explosion in the amount of images and videos that are digitally available, *e.g.* in broadcasting archives, social media sharing websites, and personal collections. The following two statistics clearly underline this observation. According to Business Insider[1] Facebook had 350 million photo uploads *per day* in 2013. The world leader in internet infrastructure Cisco estimates that "Globally, IP video traffic will be 80% of all IP traffic (both business and consumer) by 2019, up from 67% in 2014." (cis, 2015). These unprecedented large quantities of visual data motivate the need for computer vision techniques to assist retrieval, annotation, and navigation of visual content.

Arguably, the ultimate goal of computer vision as a scientific and engineering discipline is to be able to build general purpose "intelligent" vision systems. Such a system should be able to "represent" (store in an internally useful format), "interpret" (map input to this format), and "understand" (infer facts about the input based on the representation) at a high semantic level the scene depicted in an image, or a dynamic scene that unfolds in a video. Let us try to clarify these desiderata by giving more concrete examples. Scene understanding involves determining which type of objects are present in a scene, where they are, how they interact with each other, *etc*. These questions require high-level semantic interpretation of the scene, which abstracts away from many of the physical geometric and photometric properties such as viewpoint, illumination, blur,

---

[1]See http://www.businessinsider.com

etc.[2] High-level scene understanding is of central interest to the computer vision research community since it supports a large variety of applications, including text-based image and video retrieval, annotation and filtering of image and video archives, surveillance, visual recommendation systems (query by image), object and event localization (possibly embedded in (semi-)autonomous vehicles and drones), *etc*.

Scene understanding can be formulated using representations at different levels of detail, which leads to different well defined tasks that are studied in the research community. Restricting the scene interpretation to the level of object categories, we can for example distinguish the following tasks. *Image categorization* gives a very coarse interpretation of the scene: the goal is to determine if an image contains one or more objects of a certain category, *e.g.* cars, or not. In essence a single bit of information is predicted for the image. In *object localization* the task is to predict the number and location of instances of the category of interest, typically by means of a tight enclosing bounding boxes of the objects. Finally, *semantic segmentation* gives the most detailed interpretation, and assigns a category label to each pixel in the image, or classifies it as background.

These three tasks have been in a way the "canonical" tasks to study scene understanding. They have been heavily studied over the last decade, and tremendous progress has since then been made. Important benchmark datasets to track this progress are the PASCAL Visual Object Classes challenge (yearly 2005–2012) (Everingham et al., 2010), and ImageNet challenge (yearly since 2010) (Deng et al., 2009). In the video domain the corresponding canonical tasks at the level of action categories are video categorization (does the video contain an action of interest), temporal localization (where are the action instances located in time), and spatio-temporal localization (each action instance is captured by a sequence of bounding boxes across its temporal extent). In the video domain there has been a rapid succession of benchmark datasets, as performance on earlier datasets saturated. The TRECVID multimedia event detection (yearly since 2010) (Over et al., 2012) and THUMOS action recognition challenges (yearly since 2013) (Jiang et al., 2014) are currently among the most important benchmarks.

The rapid progress at category-level recognition was triggered by preceding progress in instance-level recognition (recognizing the very same object under different imaging conditions) based on invariant local descriptors, *e.g.* (Schmid and Mohr, 1997; Lowe, 1999), and machine learning methods, *e.g.* (Cortes and Vapnik, 1995; Jordan, 1998). Ensembles of local invariant descriptors delivered a rich representation, robust to partial occlusion

---

[2]Modeling and understanding such physical properties has of course its own uses, *e.g.* to correct for artefacts such as blur, but can also be useful to obtain invariance to such properties to facilitate high-level interpretation. Examples include illuminant invariant color descriptors for object recognition (Khan et al., 2012), and using 3D scene geometry to constrain object detectors by expected object sizes (Hoiem et al., 2008).

and changes in viewpoint and illumination. Machine learning tools proved effective to learn the structural patterns in such ensembles of local descriptors across instances of an object and scene categories, replacing earlier manually specified rule-based systems (Ohta et al., 1978). The combination of (i) local descriptors, (ii) unsupervised learning to aggregate these into global image descriptors, and (iii) linear classifiers, has been the dominant paradigm in most of scene understanding research for almost a decade. In particular local SIFT (Lowe, 2004) and HOG (Dalal and Triggs, 2005) descriptors aggregated into bag-of-visual word histograms (Sivic and Zisserman, 2003; Csurka et al., 2004) or Fisher vectors (Perronnin and Dance, 2007), and then classified using support vector machines (Cortes and Vapnik, 1995) have proven extremely effective.

The recent widespread adoption of deep convolutional neural networks (CNNs) (LeCun et al., 1989), following the success of Krizhevsky *et al*. in the ImageNet challenge in 2012 (Krizhevsky et al., 2012), is a second important step in the same data-driven direction where supervised machine learning is used to obtain better recognition models. CNNs replace the local descriptors with a layered processing pipeline that takes the image pixels as input and maps these to the target output, *e.g.* an object category label. In contrast to the use of fixed local descriptors in previous methods, the parameters of each processing layer in the CNN can be learned from data in a coherent framework.

It is probably fair to say that machine learning has been one of the key ingredients in the tremendous progress made in the last decade on computer vision problems such as automatic object recognition and scene understanding. Given the current proliferation of ever more powerful compute hardware and large image and video collections, we expect that machine learning will continue to play a central role in computer vision. In particular we expect that hybrid techniques that combine deep neural networks, (non-parametric) hierarchical Bayesian latent variable models, and approximate inference may prove to be extremely versatile to further advance the state of the art.

1.2   Contents of this document

The following chapters give an overview of our contributions on learning visual recognition models. We organize these across three topics: the Fisher vector image representation, metric learning techniques, and learning with incomplete supervision. Each of these will be the subject of one of the following three chapters.

In Chapter 2 we give a brief introduction to the Fisher vector representation, which aggregates local descriptors into a high dimensional vector of local first and second order statistics. Our contributions in this area in-

clude extensions based on modeling inter-dependencies among local image descriptors (Cinbis et al., 2012, 2016a), and spatial layout information respectively (Krapac et al., 2011). We present an approximate normalization scheme which speed-up applications for object and action localization (Oneata et al., 2014b), and discuss an application to object localization in which we weight the contribution of local descriptors based on approximate segmentation masks (Cinbis et al., 2013).

In Chapter 3 we consider metric learning techniques, which learn a task dependent distance metric that can be used to compare images of objects or scenes based on supervised training data. Our contributions include an approach to learn Mahalanobis metrics using logistic discriminant classifiers, and a non-parametric method based on nearest neighbors (Guillaumin et al., 2009b). We present a nearest-neighbor based image annotation method that learns weights over neighbors, and effectively determines the number of neighbors to use (Guillaumin et al., 2009a). We also present an image classification method based on metric learning for the nearest class-mean classifier that can efficiently generalize to new classes (Mensink et al., 2012, 2013b).

The third topic, presented in Chapter 4, is related to learning models from incomplete supervision. These include an image re-ranking model that can be applied to new queries not seen at training time(Krapac et al., 2010), and a semi-supervised image classification approach that leverages user provided tags that are only available at training time (Guillaumin et al., 2010a). Other contributions are related to the problem of associating names and faces in captioned news images (Guillaumin et al., 2008; Mensink and Verbeek, 2008; Guillaumin et al., 2012, 2010b; Cinbis et al., 2011), and learning semantic image segmentation models from partially-labeled training images or image-wide labels only (Verbeek and Triggs, 2007, 2008). For interactive image annotation we developed a method that models dependencies across different image labels, which improves predictions and helps to identify the most informative user input (Mensink et al., 2011, 2013a). We present a multi-fold multiple instance learning method to improve the learning of object localization models from training images where we only know if the object is present in the image or not (Cinbis et al., 2014).

Chapter 5 summarizes the contributions, and presents several directions for future research. A curriculum vitae with a list of patents and publications is included in Appendix A. All of my publications are publicly available online via my webpage.[3] Estimates of the number of citations (total 5493) and h-index (34) can be obtained from Google Scholar.[4]

---

[3]http://lear.inrialpes.fr/~verbeek
[4] http://scholar.google.com/citations?hl=en&user=oZGA-rAAAAAJ

Acknowledgement

The material presented here is by no means the result of only my own work. Over the years I have had the pleasure to work with excellent colleagues, and I would like to take the opportunity here to thank them all for these great collaborations. In particular I would like to thank my (former) PhD students Matthieu, Josip, Thomas, Gokberk, Dan, Shreyas, and Pauline.

Chapter 2

The Fisher vector representation: extensions and applications

The Fisher vector (FV) image representation (Perronnin and Dance, 2007), is an extension of the bag-of-visal-word (BoV) representation (Csurka et al., 2004; Leung and Malik, 2001; Sivic and Zisserman, 2003). Both representations characterize the distribution of local low-level descriptors such as SIFT (Lowe, 2004) extracted from an image. The BoV does so by using a partition of the descriptor space, and characterizing the image with a histogram that counts how many local descriptors fall into each cell of the partition. The FV extends this by also recording the mean and variance of the descriptors in each cell. This has two benefits: (i) the FV computes a more detailed representation per cell, therefore for a given representation dimensionality the FV is computationally more efficient than the BoV, and (ii) the FV is a smooth (linear and quadratic) function of the descriptors within a cell, therefore a learned classification function will inherit this smoothness which may lead to better generalization performance, as compared to a finer quantization that could be used to improve the BoV.

Contents of this chapter. In Section 2.1 we recall the Fisher kernel principle that underlies the FV, and discuss our related contributions. We present two contributions in more detail. In Section 2.2 we present an extension of the generative model underlying the FV to account for the dependencies among local image descriptors, which explains the effectiveness of the power normalization of the FV. In Section 2.3 we present approximate versions of the power and ℓ_2 normalization. This approximation is useful for object and action localization, where classification scores need to be evaluated over many candidate detection windows. Using the approximation these can be efficiently computed using integral images. Section 2.4 con-

cludes this chapter with a summary and some perspectives.

2.1 The Fisher vector image representation

The main idea of the Fisher kernel principle (Jaakkola and Haussler, 1999) is to use a generative probabilistic model to obtain a vectorial data representation of non-vectorial data. Examples of such data include time-series of varying lengths, or sets of vectors. Using generative models for such data with a finite set of parameters, the data is represented by the gradient of the log-likelihood of the data w.r.t. the model parameters.

More formally, let $X \in \mathcal{X}$ be an element of a space $\mathcal{X}$, and $p(X|\boldsymbol{\theta})$ be a probability distribution or density over this space, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_H)^\top$ is a vector that contains all H parameters of the probabilistic model. We then define the Fisher score vector of X w.r.t. $\boldsymbol{\theta}$ as the gradient of the log-likelihood of X w.r.t. the model parameters: $G_{\boldsymbol{\theta}}^X \equiv \nabla_{\boldsymbol{\theta}} \ln p(X)$. Clearly, $G_{\boldsymbol{\theta}}^X \in \mathbb{R}^H$ provides a finite dimensional vectorial representation of X, which essentially encodes in which way the parameters of the model should change in order to better fit the data X that should be encoded.

It is easy to see that the Fisher score vector depends on the parametrization of the model. For example, if we define $\boldsymbol{\theta}' = 2\boldsymbol{\theta}$ then $G_{\boldsymbol{\theta}'}^X = 2G_{\boldsymbol{\theta}}^X$. The dot-product between Fisher score vectors can be made invariant for general invertible re-parametrization by normalizing it with the inverse Fisher information matrix (FIM) (Jaakkola and Haussler, 1999). The normalized dot-product $G_{\boldsymbol{\theta}}^{X^\top} F_{\boldsymbol{\theta}}^{-1} G_{\boldsymbol{\theta}}^Y$ is referred to as the Fisher kernel. Since $F_{\boldsymbol{\theta}}$ is positive definite, we can decompose its inverse as $F_{\boldsymbol{\theta}}^{-1} = L_\theta^\top L_\theta$, and write the Fisher kernel as dot-product between normalized score vectors $\mathcal{G}_\theta^X = L_\theta G_\theta^X$. The normalized score vectors are referred to as Fisher vectors.

Perronnin and Dance (Perronnin and Dance, 2007) used the Fisher kernel principle to derive and image representation based on an i.i.d. Gaussian mixture model (GMM) over local image descriptors, such as SIFT (Lowe, 2004). In this case $X = \{x_1, \ldots, x_N\}$ is a set of N local descriptors $x_n \in \mathbb{R}^D$. The FV is given by the concatenation of the normalized gradients w.r.t. the mixing weights π_k, means μ_k, and standard deviations σ_k that characterize the GMM:

$$\mathcal{G}_{\alpha_k}^X = \frac{1}{\sqrt{\pi_k}} \sum_{n=1}^N (q_{nk} - \pi_k), \tag{2.1}$$

$$\mathcal{G}_{\mu_k}^X = \frac{1}{\sqrt{\pi_k}} \sum_{n=1}^N q_{nk} \left(\frac{x_n - \mu_k}{\sigma_k} \right), \tag{2.2}$$

$$\mathcal{G}_{\sigma_k}^X = \frac{1}{\sqrt{\pi_k}} \sum_{n=1}^N q_{nk} \frac{1}{\sqrt{2}} \left(\frac{(x_n - \mu_k)^2}{\sigma_k^2} - 1 \right), \tag{2.3}$$

where $q_{nk} = \frac{\pi_k \mathcal{N}(x_n; \mu_k, \sigma_k)}{p(x_n)}$ denotes the posterior probability that x_n was generated by the k-th mixture component. Equation (2.1) and (2.3) apply in the one-dimensional case, but also per-dimension in the multidimensional case if the Gaussian covariance matrices are diagonal.

The FV extends the bag-of-visual-words (BoV) image representation (Csurka et al., 2004; Leung and Malik, 2001; Sivic and Zisserman, 2003), which was the dominant image representation for image classification, retrieval, and object detection over the last decade. The components of the FV capture the zero, first, and second order moment of the data associated with each Gaussian component. The zero-order statistics in Eq. (2.1) can be seen as a normalized version of the soft-assign BoV representation (van Gemert et al., 2010). The normalization ensures that the representation has zero-mean and unit covariance. We refer to (Sánchez et al., 2013) for a more detailed presentation, and comparisons to other recent image representations. This paper we also include a detailed derivation of the diagonal approximation of the FIM for the Gaussian mixture case, which is particularly interesting for the mixing weights.

Perronnin *et al*. (Perronnin et al., 2010b) proposed two normalizations to improve the performance of the FV image representation. First, the power normalization consists in taking a signed power, $z \leftarrow \text{sign}(z)\text{abs}(z)^\rho$, on each dimension separately, where typically $\rho = 1/2$. Second, the ℓ_2 normalization scales the FV to have unit ℓ_2 norm. The power normalization leads to a discounting of the effect of large values in the FV. This is useful to counter the burstiness effect of local visual descriptors, which is due to the locally repetitive structure of images. Winn *et al*. (Winn et al., 2005) applied square-root transformation to model BoV histograms in a generative classification model motivated as a variance stabilizing transformation. Jégou *et al*. (Jégou et al., 2009) applied square-root transformation to BoV histograms to counter burstiness and improve image retrieval performance. Similarly, the square-root transform has also proven to be effective to normalize histogram-based SIFT features SIFT (Arandjelović and Zisserman, 2012) which exhibit similar burstiness effects. The power normalization has also been applied to VLAD representations (Jégou et al., 2012), which is a simplified version of the FV based on k-means instead of GMM clustering, and only uses the first-order statistics of the assigned descriptors. In (Kobayashi, 2014) Kobayashi models BoV histograms and SIFT descriptors using Dirichlet (mixture) distributions, which yields logarithmic transformations with a similar effect as power normalization.

In (Cinbis et al., 2012) we address the burstiness effect in a different manner. Our observation is that the GMM and multinomial model over local descriptors that underlie the FV and BoV are i.i.d., which is does not reflect the (locally) repetitive nature of local image descriptors. We therefore define models in which the local descriptors are no longer i.i.d. By treat-

ing the parameters of the original generative models as latent variables, we render the local descriptors mutually dependent. This builds a burstiness effect in the data that is sampled from the model. We show that the FV of such non-i.i.d. models naturally exhibit similar discounting effects as otherwise obtained using power normalization. Experimentally, we also observe similar performance improvements as obtained using power normalization. We present this work in more detail in Section 2.2.

Localization of objects in images, and actions in video, is often formulated as a large-scale classification problem, where many possible detection regions are scores, and the region with maximum response is retained (Dalal and Triggs, 2005; Felzenszwalb et al., 2010). Efficient localization techniques often rely on the additivity of the region representation over local descriptors (Chen et al., 2013a; Lampert et al., 2009a; Viola and Jones, 2004). For example, when combining additive representations with linear score functions, scores can be computed per local descriptor and integrated over arbitrarily large regions in constant time using integral images (Chen et al., 2013a). While power and ℓ_2 normalization improve the performance of the FV representation, they make the representation non-additive over the local descriptors. In (Oneata et al., 2014b) we present approximate versions of the power and ℓ_2 normalization which allow us to efficiently compute linear score functions of the normalized FV. The approximations allow the use of integral images to efficiently compute sums of scores, assignments, and norms of local descriptors per visual word. We also show how our approximations can be used in branch-and-bound search (Lampert et al., 2009a) to further speed-up the localization process. Experimentally we find that the approximation have only a limited impact on the localization performance, but lead to more than an order of magnitude speed-up. We present this work in more detail in Section 2.3. Although not experimentally explored, this approach can also be used in combination with our supervoxel-based spatio-temporal detection proposal method presented in (Oneata et al., 2014a). In that case, however, integral images cannot be used due to the irregular supervoxel structure.

The FV and most other local image descriptor aggregation methods like BoV and VLAD, are invariant for the spatial arrangement of local image descriptors. This invariance is beneficial in the sense of making the representation robust, *e.g.* to deformation of articulated objects, or re-arrangement of objects in a scene. In certain cases some degree of spatial layout information is useful however, *e.g.* to accurately localize objects in a scene (no effect of re-arrangements) (Cinbis et al., 2013), or to recognize rigid objects (no effects of articulation) (Simonyan et al., 2013). The spatial pyramid (SPM) approach (Lazebnik et al., 2006) is one of the most basic methods to capture spatial layout. It concatenates representations of several image regions at different positions and scales. The disadvantage of this approach —in particular for high dimensional representation like the FV— is that the size

Figure 2.1 – Segmentation masks for two detection windows. The first three columns show the window, our weighted mask, and the masked window. The eight images on the right show the individual binary masks of superpixels lying fully inside the window, for each of eight segmentations.

of the representation grows linearly with the number of regions. In (Krapac et al., 2011) we proposed an alternative approach where we instead model layout using a "spatial FV" over the 2D spatial positions of the local descriptors assigned to each visual word. Since the local descriptors are typically higher dimensional, *e.g.* 128 dim. for SIFT, modeling the 2D spatial coordinates increases the representation size only marginally, as opposed to the SPM which multiplies the representation size by the number of cells. Sánchez *et al*. (Sánchez et al., 2012) developed a related approach, which consists in appending the position coordinates to local descriptors, and encoding these with a usual FV representation. The spatial FV and SPM are complementary techniques that can be combined by concatenating the spatial FV representation computed over several image regions. In (Wang et al., 2015) we found this combination to be most effective to encode the layout of local spatio-temporal features (Wang and Schmid, 2013) for action recognition and localization in video.

In (Cinbis et al., 2013) we presented a refined FV representation which reduces the detrimental effect of background clutter to improve object localization. We use an approximate segmentation mask with which we weight the contribution of local descriptors in the FV: each term in equations (2.1)—(2.3) is multiplied by the corresponding value in the mask. To compute our masks we rely on superpixels, which tend to align with object boundaries. If superpixels traverse the window boundary, then it is likely to be either part of a background object that enters into the detection window, or to be a part of an object of interest which extends outside the window. In both cases we would like to suppress such regions, either because it introduces clutter, or because the window is to small w.r.t. the object. Based on this observation, we compute a binary segmentation mask for a detection window, by masking out any superpixel that is not fully inside the detection window. Since we cannot expect the superpixel segmentation to perfectly align with object boundaries, we compute a weighted segmentation mask by averaging over binary masks obtained using super-

pixels of several granularities, and based on different color channels. The way we derive our masks is related to the superpixel straddling score that that was used in (Alexe et al., 2012) to find high-recall candidate detection windows for generic object categories. See Figure 2.1 for an illustration of these masks.

Associated publications. Here, we list the most important publications associated with the contributions presented in this chapter, together with the number of citations they have received.

- (Cinbis et al., 2016a) G. Cinbis, J. Verbeek, C. Schmid. *Approximate Fisher kernels of non-iid image models for image categorization.* IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear, 2015. Citations: 2

- (Wang et al., 2015) H. Wang, D. Oneaţă, J. Verbeek, C. Schmid. *A robust and efficient video representation for action recognition.* International Journal of Computer Vision, to appear, 2015. Citations: 7

- (Sánchez et al., 2013) J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek. *Image classification with the Fisher vector: theory and practice.* International Journal of Computer Vision 105 (3), pp. 222–245, 2013. Citations: 332

- (Oneata et al., 2014b) D. Oneaţă, J. Verbeek, C. Schmid. *Efficient Action Localization with Approximately Normalized Fisher Vectors.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2014. Citations: 18

- (Cinbis et al., 2013) G. Cinbis, J. Verbeek, C. Schmid. *Segmentation Driven Object Detection with Fisher Vectors.* Proceedings IEEE International Conference on Computer Vision, December 2013. Citations: 63

- (Oneata et al., 2013) D. Oneaţă, J. Verbeek, C. Schmid. *Action and Event Recognition with Fisher Vectors on a Compact Feature Set.* Proceedings IEEE International Conference on Computer Vision, December 2013. Citations: 132

- (Cinbis et al., 2012) G. Cinbis, J. Verbeek, C. Schmid. *Image categorization using Fisher kernels of non-iid image models.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2012. Citations: 36

- (Krapac et al., 2011) J. Krapac, J. Verbeek, F. Jurie. *Modeling spatial layout with Fisher vectors for image categorization.* Proceedings IEEE International Conference on Computer Vision, November 2011. Citations: 124

2.2 Modeling local descriptor dependencies

The use of non-linear feature transformations in bag-of-visualword (BoV) histograms has been widely recognized to be beneficial for image categorization. Popular examples include the use of chi-square kernels (Leung and Malik, 2001; Zhang et al., 2007), or taking the square-root of histogram entries (Perronnin et al., 2010a,b), also referred to as the Hellinger kernel (Vedaldi and Zisserman, 2010). The effect of these is similar. Both transform the features such that the first few occurrences of visual words will have a more pronounced effect on the classifier score than if the count is increased by the same amount but starting at a larger value. This is desirable, since now the first patches providing evidence for an object category can significantly impact the score, *e.g.* making it easier to detect small objects.

In this section we will re-consider the i.i.d. assumption that underlies the FV image representation (Perronnin and Dance, 2007; Sánchez et al., 2013). In particular we consider exchangeable models that treat the parameters of the i.i.d. models as latent variables, and integrate these out to obtain a non-i.i.d. model. It turns out that non-linear feature transformations similar to those that have been found effective in the past arise naturally from our latent variable models. This suggests that such transformations are successful *because* they correspond to a more realistic non-i.i.d. model.

More technical details and experimental results can be found in the original CVPR'12 paper (Cinbis et al., 2012) and the forthcoming extended PAMI paper (Cinbis et al., 2016a). An electronic version of the latter is available at https://hal.inria.fr/hal-01211201/file/paper.pdf

2.2.1 Interpreting the BoV representation as a Fisher vector

We will first re-interpret the popular BoV representation as a FV of a simple multinomial model over the visual words extracted from an image. Let us use $w_{1:N} = \{w_1, \ldots, w_N\}$, with $w_n \in \{1, \ldots, K\}$, to denote the set of discrete visual word indices assigned to the N local descriptors extracted from an image. We model $w_{1:N}$ as being i.i.d. distributed according to a

Figure 2.2 – The visible image patches are assumed to be uninformative on the masked ones by the independence assumption. Clearly, local image patches are *not* i.i.d.: one can predict with high confidence the appearance of the hidden image patches from the visible ones.

multinomial distribution:

$$p(w_{1:N}) = \prod_{n=1}^{N} p(w_n) = \prod_{n=1}^{N} \pi_{w_n}, \tag{2.4}$$

$$\pi_k = \frac{\exp(\alpha_k)}{\sum_{k'=1}^{K} \exp(\alpha_{k'})}. \tag{2.5}$$

The k-th element of the Fisher score vector for this model then equals:

$$\frac{\partial \ln p(w_{1:N})}{\partial \alpha_k} = \sum_{n=1}^{N} [\![w_n = k]\!] - N\pi_k, \tag{2.6}$$

where $[\![\cdot]\!]$ is the Iverson bracket notation that equals one if the expression in its argument is true, and zero otherwise. The first term counts the number of occurrences of visual word k. Concatenating the partial derivatives we obtain the Fisher score vector as $\nabla_\alpha \ln p(w_{1:N}) = h - N\pi$, where $h \in \mathbb{R}^K$ is the histogram of visual word counts, and $\pi \in \mathbb{R}^K$ is the vector of the multinomial probabilities. Note that this is just a shifted version of the visual word histogram h, which centers the representation at zero; the constant shift by $N\pi$ is irrelevant for most classifiers.

The sum in Eq. (2.6), and therefore the observed histogram form, is an immediate consequence of the i.i.d. assumption in the model. To underline the boldness of this assumption, consider Figure 2.2 where visible image patches are assumed to be uninformative on the masked ones by the independence assumption.

2.2.2 A non-i.i.d. BoV model

We will now define an alternative non-i.i.d. model for visual word indices, which maintains exchangeability among the variables, *i.e.* the ordering of the visual word indices is irrelevant as in the i.i.d. model. To this end, we define the multinomial π to be a latent variable per image, and drawn the visual word indices i.i.d. from this mulitinomial. This construction ties
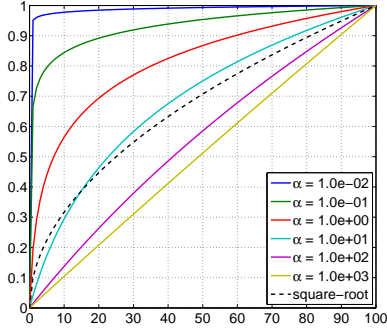
Figure 2.3 – Digamma functions $\psi(\alpha + h)$, for various α, and $\sqrt{h}$ as a function of n. All functions have been re-scaled to the range $[0, 1]$.

all visual word indices together, since knowing some visual word indices gives information on the unknown π, which in turn influences predictions on other visual word indices. We assume a conjugate Dirichlet prior distribution over the multinomial π. Formally, this model is defined as

$$p(\pi) = \mathcal{D}(\pi|\alpha), \tag{2.7}$$

$$p(w_{1:N}) = \int_{\pi} p(\pi) \prod_{n=1}^{N} p(w_n|\pi) = \frac{\Gamma(\hat{\alpha})}{\Gamma(N + \hat{\alpha})} \prod_{k} \frac{\Gamma(h_k + \alpha_k)}{\Gamma(\alpha_k)}, \tag{2.8}$$

where $\Gamma(\cdot)$ is the Gamma function, $\hat{\alpha} = \sum_k \alpha_k$, and h_k is the count of visual word k among $w_{1:N}$. This model is known as the compound Dirichlet-multinomial distribution, or multivariate Pólya distribution.

To better understand the dependency structure implied by this model, it is instructive to consider the conditional probability of a new index given a number of preceding indices:

$$p(w = k|w_{1:N}) = \int_{\pi} p(w = k)p(\pi|w_{1:N}) = \frac{h_k + \alpha_k}{N + \hat{\alpha}}. \tag{2.9}$$

The model predicts an index k with probability proportional α_k plus its count h_k among preceding indices. Therefore, the smaller the α_k are, the stronger the conditional dependence becomes.

The partial derivative of the log-likelihood of the model w.r.t. α_k is

$$\frac{\partial \ln p(w_{1:N})}{\partial \alpha_k} = \psi(\alpha_k + n_k) + \text{const.} \tag{2.10}$$

where $\psi(x) = \partial \ln \Gamma(x)/\partial x$ is the digamma function, and the constant does not depend on $w_{1:N}$. Therefore, the Fisher score is determined by $\psi(\alpha_k + h_k)$ up to additive constants, *i.e.* it is given by a transformation of the visual word counts n_k. Figure 2.3 shows the transformation $\psi(\alpha + h)$ for various values of α, along with the square-root function for reference. We see that, depending on the value of α, the digamma function produces a qualitatively similar monotone-concave transformation of the histogram entries as the square-root.

2.2.3 Extension to GMM data models

The same principle, that we used above to obtain an exchangeable non-i.i.d. model on the basis of a multinomial model, can also be applied the i.i.d. GMM data model that is typically used in FV representations. We again treat the model parameters as latent variables and place conjugate priors on the GMM parameters: a Dirichlet prior on the mixing weights, and a combined Normal-Gamma prior on the means μ_k and precisions $\lambda_k = \sigma_k^{-1}$:

$$p(\lambda_k) = \mathcal{G}(\lambda_k|a_k, b_k), \tag{2.11}$$
$$p(\mu_k|\lambda_k) = \mathcal{N}(\mu_k|m_k, (\beta_k\lambda_k)^{-1}). \tag{2.12}$$

The distribution on the descriptors $x_{1:N}$ in an image is obtained by integrating out the latent GMM parameters:

$$p(x_{1:N}) = \int_{\pi,\mu,\lambda} p(\pi)p(\mu,\lambda)\prod_{i=1}^{N} p(x_i|\pi,\mu,\lambda), \tag{2.13}$$

$$p(x_i|\pi,\mu,\lambda) = \sum_k \pi_k \mathcal{N}(x_i|\mu_k, \lambda_k^{-1}), \tag{2.14}$$

where $p(w_i = k|\pi) = \pi_k$, and $p(x_i|w_i = k, \lambda, \mu) = \mathcal{N}(x_i|\mu_k, \lambda_k^{-1})$ is the Gaussian corresponding to the k-th visual word.

Unfortunately, computing the log-likelihood in this model is intractable, and so is the computation of its gradient required for hyper-parameter learning and extracting the FV representation. To overcome this problem we propose to approximate the log-likelihood by means of a variational lower bound (Jordan et al., 1999). We optimize this bound to learn the model, and compute its gradients as an approximation to the true Fisher score for this model. Our use of variational free-energies to derive Fisher kernels differs from (Perina et al., 2009b,a), which define an alternative encoding based consisting of a vector of summands of the free-energy of a generative model.

2.2.4 Experimental validation

We validate the latent variable models proposed above with image categorization experiments using the PASCAL VOC 2007 dataset (Everingham et al., 2010). We use the standard evaluation protocol and report the mean average precision (mAP) across the 20 object categories. As a baseline, we follow the experimental setup described in evaluation study of Chatfield *et al*. (Chatfield et al., 2011). We compare global image representations, and representations that capture spatial layout by concatenating the signatures computed over eight spatial cells as in the spatial pyramid matching (SPM) method (Lazebnik et al., 2006). We use linear SVM classifiers, and we cross-validate the regularization parameter.
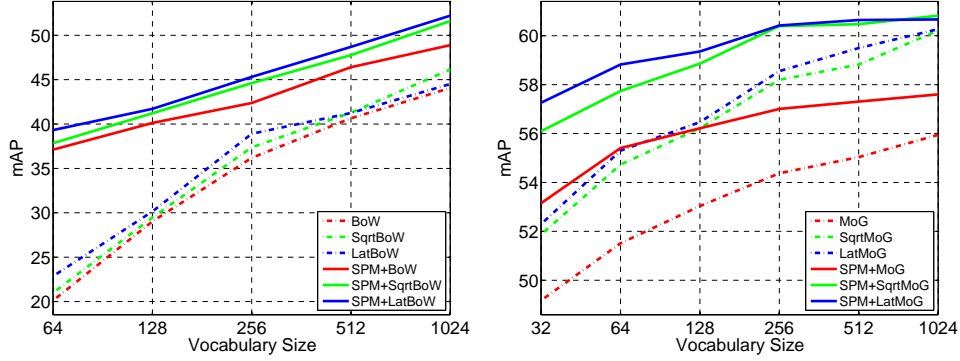
Figure 2.4 – Comparison of BoV (left) and GMM (right) representations: no transformation (red), signed square-root (green) and latent variable model (blue). With SPM (solid) and without (dashed).

Before training the classifiers we apply two normalizations to the representations. First, we whiten the representations so that each dimension is zero-mean and has unit-variance across images, this corresponds to an approximate normalization with the inverse Fisher information matrix (Krapac et al., 2011). Second, following (Perronnin et al., 2010b), we apply ℓ_2 normalization.

In the left panel of Figure 2.4 we compare the results obtained using standard BoV histograms, square-rooted histograms, and the Pólya model. Overall, we see that the spatial information of SPM is useful, and that larger vocabularies increase performance. We observe that square-rooting and the Pólya model both consistently improve the BoW representation. Furthermore, the Pólya model generally leads to larger improvements than square-rooting. These results confirm the observation made above that the non-i.i.d. Pólya model generates similar transformations on BoW histograms as square-rooting does, providing a model-based explanation of why square-rooting is beneficial.

In the right panel of Figure 2.4 we compare image representations based on Fisher vectors computed over GMM models, their square-rooted version, and the latent GMM model. We can observe that the GMM representations lead to better performance than the BoV ones while using smaller vocabularies. Furthermore, the discounting effect of our latent model and square rooting has a much more pronounced effect here than it has for BoV models, improving mAP scores by around 4 points. Also here our latent models lead to improvements that are comparable and often better than those obtained by square-rooting. So again, the benefits of square-rooting can be explained by using non-i.i.d. latent variable models that generate similar representations.

2.2.5 Summary

We have presented latent variable models for local image descriptors, which avoid the common but unrealistic i.i.d. assumption. The Fisher vectors of our non-i.i.d. models are functions computed from the same sufficient statistics as those used to compute Fisher vectors of the corresponding i.i.d. models. These functions are similar to transformations that have been used in earlier work in an ad hoc manner, such as the power normalization, or signed-square-root. Our models provide an explanation of the success of such transformations, since we derive them here by removing the unrealistic i.i.d. assumption from the popular BoW and MoG models. The Fisher vectors for the proposed intractable latent MoG model can be successfully approximated using the variational Fisher vector framework. In (Cinbis et al., 2016a) we further show that the FV of our non-i.i.d. MoG model over CNN image region descriptors is also competitive with state-of-the-art feature aggregation representations based on i.i.d. models.

2.3 Approximate Fisher vector normalization

The recognition and localization of human actions and activities is an important topic in automatic video analysis. State-of-the-art temporal action localization (Oneata et al., 2013) is based on Fisher vector (FV) encoding of local dense trajectory features (Wang and Schmid, 2013). Recent state-of-the-art action recognition results of (Fernando et al., 2015; Peng et al., 2014) are also based on extensions of this basic approach. The power and ℓ_2 normalization of the FV, introduced in (Perronnin et al., 2010b), significantly contribute to its effectiveness. The normalization, however, also renders the representation non-additive over local descriptors. Combined with its high dimensionality, this makes the FV computationally costly when used for localization tasks. In this section we present an approximate normalization scheme, which significantly reduces the computational cost of the FV when used for localization, while only slightly compromising the performance.

For more technical details and experimental results we refer to the CVPR paper (Oneata et al., 2014b), which is available at `https://hal.inria.fr/hal-00979594/file/efficient_action_localization.pdf`

2.3.1 Efficient action localization in video

Localization of actions in video, and similarly objects in images, can be considered as a large-scale classification problem, where we want to find the highest scoring windows in a video or image w.r.t. a classification model of the category of interest. Unlike generic large-scale image classification, however, the problem is highly structured in this case, in the sense that all

windows are crops of the same video or image under consideration. This structure has been extensively exploited in the past. In particular, when the features for a detection window are obtained as sums of local features, integral images can be used to pre-compute cumulative feature sums. Once the integral images are computed, these can be used to compute the sums of local features in constant-time w.r.t. the window size. Viola and Jones (Viola and Jones, 2004) used this idea to efficiently compute Haar filters for face detection. Recently, Chen *et al.* (Chen et al., 2013a) used the same idea to aggregate scores of local features in an object detection system based on a non-normalized FV representation. Another way to exploit the structure of the localization problem is to use branch-and-bound search, as *e.g.* used by Lampert *et al.* (Lampert et al., 2009a) for object localization in images, and by Yuan *et al.* (Yuan et al., 2009) for spatio-temporal action localization in video. Instead of evaluating the score of one window at a time, they hierarchically decompose the set of detection windows and consider upper-bounds on the score of sets of windows to explore the most promising ones first. For linear classifiers, such bounds can again be efficiently computed using integral image representations.

While power and ℓ_2 normalization have proven effective to improve the performance of the FV (Oneata et al., 2013; Perronnin et al., 2010b), the resulting normalized FV is no longer additive over local features. Therefore, these FV normalizations prevent the use of integral image techniques to efficiently aggregate local features or scores when assessing larger windows. As a result, most of the recent work that uses FV representations for object and action localization, and semantic segmentation, either uses efficient — but performance-wise limited— additive non-normalized FVs (Chen et al., 2013a; Csurka and Perronnin, 2011) or explicitly computes normalized FVs for all considered windows (Cinbis et al., 2013; Oneata et al., 2013). The recent work of Li *et al.* (Li et al., 2013) is an exception to this trend; they present an efficient approach to incorporate exact ℓ_2 normalization. Their approach, however does not provide an efficient approach to incorporate the power-normalization, which they therefore only apply locally.

Approximate power normalization. In (Cinbis et al., 2012), see Section 2.2, we have argued that the power normalization corrects for the independence assumption that is made in the GMM model that underpins the FV representation. We presented latent variable models which do not make this independence assumption, and experimentally found that such models lead to similar performance improvements as the power-normalization. In particular, we showed that the gradients w.r.t. the mixing weights in the non-i.i.d. model take the form a BoV histogram transformed by the digamma function, which —like the power-normalization— is concave and monotonically increasing function. The components of the FV of the non-

i.i.d. model corresponding to the means and variances can also be shown to be related to the FV of the i.i.d. model by a monotone concave function that is constant per visual word. Based on this analysis, we propose and approximate version of the power normalization.

Recall that the components of the FV that correspond to the gradients w.r.t. the means and variances take the form of *weighted sums*, see equations (2.2) and (2.3). Let us write these in a more compact and abstract manner as:

$$G_k = \sum_n q_{nk} g_{nk} = \left(\sum_n q_{nk} \right) \sum_n \frac{q_{nk} g_{nk}}{\sum_m q_{mk}}, \qquad (2.15)$$

where q_{nk} and g_{nk} denote the weight and gradient contribution of the n-th local descriptor for the k-th Gaussian. The right-most form in Eq. (2.15) re-interprets the FV as a *weighted average* of local contributions, multiplied by the sum of the weights. The power-normalization is computed as an element-wise signed-power of G_k. In our approximation we, instead, apply the power only to the positive scalar given by the sum of weights:

$$\mathcal{G}_k = \left(\sum_n q_{nk} \right)^\rho \sum_n \frac{q_{nk} g_{nk}}{\sum_m q_{mk}}. \qquad (2.16)$$

Our approximate power normalization does not affect the orientation of the FV, but only modifies its magnitude, which grows sub-linearly with the sum of weights to account for the burtiness of local descriptors.

We concatenate the $\mathcal{G}_k$ for all Gaussians to form the normalized FV $\mathcal{G} = [\mathcal{G}_1, \ldots, \mathcal{G}_K]$. Using our approximate power-normalization, a linear (classification) function can be computed by aggregating local scores. For a weight vector $w = [w_1, \ldots, w_K]$ we have:

$$\langle w, \mathcal{G} \rangle = \sum_k \left(\sum_n q_{nk} \right)^\rho \sum_n \frac{q_{nk} \langle w_k, g_{nk} \rangle}{\sum_n q_{nk}} \qquad (2.17)$$

$$= \sum_k \left(\sum_n q_{nk} \right)^{\rho-1} \sum_n s_{nk}, \qquad (2.18)$$

where the $s_{nk} = q_{nk} \langle w_k, g_{nk} \rangle$ denote the scores of the local non-normalized FV. These scores can be pre-computed, and added over detection windows in constant time using integral images.

Approximate ℓ_2 normalization. We now proceed with an approximation of the ℓ_2 norm of $\mathcal{G}$. The squared ℓ_2 norm is a sum of squared norms per Gaussian component: $||\mathcal{G}||_2^2 = \sum_k \mathcal{G}_k^\top \mathcal{G}_k$. From Eq. (2.16) we have

$$\mathcal{G}_k^\top \mathcal{G}_k = \left(\sum_n q_{nk} \right)^{2(\rho-1)} \sum_{n,m} q_{nk} q_{mk} \langle g_{nk}, g_{mk} \rangle. \qquad (2.19)$$
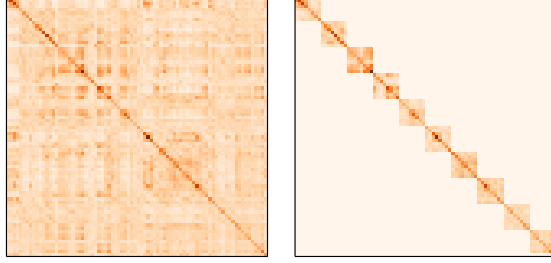
Figure 2.5 – Visualization of dot-products between frame-level FVs summed in Eq. (2.19) (left). Most large values lie near the diagonal due to local temporal self-similarity, which motivates a block diagonal approximation (right).

We approximate the double sum over dot products of local gradient contributions by assuming that most of the local gradients will be near orthogonal for high-dimensional FVs. This leads to an approximation $L(\mathcal{G}_k)$ of the squared ℓ_2 norm of $\mathcal{G}_k$ computed from sums of local quantities:

$$L(\mathcal{G}_k) = \left(\sum_n q_{nk} \right)^{2(\rho-1)} \sum_n q_{nk}^2 l_{nk}, \qquad (2.20)$$

where $l_{nk} = \langle g_{nk}, g_{nk} \rangle$ is the local squared ℓ_2 norm. Summing these over the visual words, we approximate $\|\mathcal{G}\|_2^2$ with $L(\mathcal{G}) = \sum_k L(\mathcal{G}_k)$.

Figure 2.5 visualizes for a typical video the dot-products between frame-level FVs g_{nk}; where the frame-level FVs are computed using Eq. (2.15). Instead of dropping all off-diagonal terms, we can make a block-diagonal approximation by first aggregating the frame-level descriptors over several frames, and using these the local FVs. In particular, if for action localization we use a temporal stride of s frames, then we aggregate local features across blocks of s frames into a single FV.

We now combine the above approximations to compute a linear function of our approximately normalized FV as

$$f(\mathcal{G}; w) = \left\langle w, \mathcal{G}/\sqrt{L(\mathcal{G})} \right\rangle = \langle w, \mathcal{G} \rangle \Big/ \sqrt{L(\mathcal{G})}. \qquad (2.21)$$

To efficiently compute $f(\mathcal{G}; w)$ over many windows of various sizes and positions, we can use integral images. We need to compute $3K$ integral images: one for the assignments q_{nk}, scores s_{nk}, and norms l_{nk} of each visual word. The cost to compute the integral images is $O(Kd)$, for K Gaussian components and d dimensional local descriptors. Using these integral images, the cost to score an arbitrarily large window is $O(K)$. In comparison, when using exact normalization we need to compute $2Kd$ integral images, which costs $O(Kd)$, after which we can score arbitrarily large windows at a cost $O(Kd)$. Thus our approximation leads to the following advantages: (i) it requires us to compute and store a factor $2d/3$ less integral images (but the computational complexity is the same), and (ii) it allows us to score windows with an $O(d)$ speed-up, once the integral images are computed.

Integration with branch-and-bound search Our approximations can be used to speed-up sliding window localization for actions in video, or for objects in still images. Our approximations can also be used for localization with branch-and-bound search instead of exhaustive sliding window search. We follow the approach of Lampert *et al*. (Lampert et al., 2009a) to structure the search space into sets of windows by defining intervals for each of the boundaries of the search window, and branching the space by splitting these intervals. We can derive upperbounds on linear score functions of the approximately normalized FV for such sets of windows. These bounds can be efficiently evaluated using integral images over the scores, weights, and norms of the local FVs. For sake of brevity we do not present them here, and refer to (Oneata et al., 2014b) instead.

2.3.2 Experimental evaluation

We present results of action localization experiments to evaluate the impact of our approximate FV normalizations on localization accuracy and speed.[1] In our experiments we use the common setting of $\rho = \frac{1}{2}$, see *e.g.* (Chatfield et al., 2011; Sánchez et al., 2013), which corresponds to a signed square-root.

   We use two datasets extracted from feature lenght movies. The *Coffee and Cigarettes* dataset (Laptev and Pérez, 2007) is annotated with instances of two actions: *drinking* and *smoking*. The Duchene dataset (Duchenne et al., 2009) is annotated with the actions *open door* and *sit down*. To evaluate localization we follow the standard protocol (Duchenne et al., 2009; Laptev and Pérez, 2007), and report the average precision (AP), using a 20% intersection-over-union threshold. For localization we consider a sliding temporal window approach with lengths from 20 to 180 frames, with increments of 5 frames. We use a stride of five frames to locate the windows on the video. As in (Oneata et al., 2013), we use zero-overlap non-maximum suppression, and re-scale the window scores by the duration.

   We use the dense trajectory features of Wang *et al*. (Wang et al., 2013), and encode them in a 16K dimensional FV using a GMM with $K = 128$ components and MBH features projected to $d = 64$ dimensions with PCA. We use linear SVM classifiers for our detectors, and cross-validate the regularization parameter and the class balancing weight.

   In Table 2.1 we assess the effect of exact and approximate normalization in terms of localization performance and speed. For all four actions the power and ℓ_2 normalization improve the results dramatically, improving the mean AP from 16.4% to 41.9%. This improvement, however, comes at a 64 fold increase in the computation time. Using our approximate normalization we obtain a mean AP of 37.7%, which is relatively close to the

---

[1]Results are taken from (Oneata, 2015), which differ from those in (Oneata et al., 2014b) in the used features, but include results for the Duchenne dataset (Duchenne et al., 2009) not reported in (Oneata et al., 2014b).

Normalization	Drinking	Smoking	Open Door	Sit Down	mean AP	Speed-up
None	34.0	15.6	10.3	5.9	16.4	64×
Approximate	67.1	52.0	18.1	13.6	37.7	16×
Exact	64.8	55.4	28.4	19.0	41.9	1×

Table 2.1 – Action localization performance using either no, exact, or approximate normalization.

41.9% using exact normalization, while being 16 times faster to compute that exact normalization.

2.3.3   Summary

We have presented approximate versions of the power and ℓ_2 normalization of the Fisher vector representation. These approximations allow efficient evaluation of linear score functions for localization applications, by caching local per visual word sums of scores, assignments, and norms. In (Oneata et al., 2014b) we also derive efficient bounds on the score that permit the use of our approximations in branch-and-bound search. Experimental results for action classification and localization show that these approximations only have a limited impact on performance, while yielding speedups of at least one order of magnitude.

The efficient localization techniques presented here are directly applicable to other localization tasks, such as object localization in still images, and spatio-temporal action localization. Since these tasks consider higher dimensional search spaces, we expect the speedup of our approximations, as well as branch-and-bound search, to be even larger than for temporal localization task that we considered in this paper.

2.4   Summary and outlook

This chapter presented our contributions related to the Fisher vector image representation, and highlighted two contributions. The first derives a representation based on exchangeable non-iid models, which gives rise to discounting effects that are usually ensured via transformations such as power-normalization. The second contribution is an approximate normalization scheme that allows significant speedups when using Fisher vectors for localization tasks.

While recently CNNs have replaced methods based on local features and FV-pooling in state-of-the-art object recognition and detection systems, we believe that the Fisher kernel will remain a relevant technique. First, in domains where training data is scarce (*e.g.* using imagery from a-typical

spectral bands such as infra-red, or in unusual conditions such as submarine imagery), it might not be feasible to effectively learn deep architectures with millions of parameters (due to the lack of data to even pretrain the model). Second, FV-type feature pooling can be used as a component of end-to-end trainable CNNs as an alternative or in addition to the commonly used max-pooling, see *e.g.* (Arandjelović et al., 2015). Third, the Fisher kernel principle may prove useful to derive representations from powerful deep generative latent variable image models (Gregor et al., 2015), which can be trained with little or no supervision.

Chapter 3

Metric learning approaches for visual recognition

Notions of similarity or distance to compare images, videos, or fragments of these, are pervasive in computer vision problems. Examples include comparing local image descriptors (*e.g.* for dictionary learning), computing distances among full-image descriptors (*e.g.* for image retrieval), and specific object descriptors (*e.g.* for face verification: are two face images of the same person or not?). More indirect examples include nearest neighbor classification to propagate annotations from training examples to new visual content, and the use of distances to define contrast sensitive pairwise potentials in vision problems that are cast as optimization problems in random fields. Metric learning techniques are used to acquire measures of similarity or distances to compare images or other objects, based on supervised training data. By learning the metric from representative training data, a problem specific metric can be learned which is generally more effective, since it can be trained to ignore irrelevant features and emphasize others.

Contents of this chapter. In Section 3.1 we give an overview of our contributions in this area in the context of related work in the literature. After that, we present two contributions in more detail. In Section 3.2 we present a nearest neighbor image annotation method that annotates new images by propagating the annotation keywords of the most similar training images. We use a probabilistic formulation to learn the weights by which the nearest neighbors are taken into account. In Section 3.3 we consider learning of metrics for nearest-mean classifiers. Such classifiers are attractive in settings where images of new and existing classes arrive continuously, since they only require computing the mean of the image signatures associated with a class. In Section 3.4 we briefly summarize the contributions from this chapter.

24

3.1 Contributions and related work

One of the most prevalent forms of metric learning aims to find Mahalanobis metrics. These metrics generalize the Euclidean distance, and take the form $d_M(x_i, x_j) = (x_i - x_j)^\top M(x_i - x_j)$, where M is a positive definite matrix, which can be decomposed as $M = L^\top L$. Due to this decomposition, we can write the Mahalanobis distance in terms of L as: $d_M(x_i, x_j) = ||L(x_i - x_j)||_2^2$. Which shows that we can interpret the Mahalanobis distance as the squared Euclidean distance after a linear transformation of the data. Most supervised Mahalanobis metric learning methods are based on loss functions defined over pairs or triplets of data points, see *e.g.* (Davis et al., 2007; Globerson and Roweis, 2006; Guillaumin et al., 2009b; Köstinger et al., 2012; Mignon and Jurie, 2012; Wang et al., 2014b; Weinberger and Saul, 2009). We refer the reader to recent survey papers (Bellet et al., 2013; Kulis, 2012) for a detailed review of these. Methods based on pairwise loss terms, such as *e.g.* (Davis et al., 2007), learn a metric so that positive pairs (*e.g.* points having the same class label) have a distance that is smaller than negative pairs (*e.g.* points with different class labels). Triplet-based approaches, such as LMNN (Weinberger and Saul, 2009), do not require that all distances between positive pairs are smaller than those between negative pairs. Instead, they consider triplets, where x_i is an 'anchor point' for which the nearest points from the same class should be closer than any points form different classes.

In (Guillaumin et al., 2009b) we presented two metric learning methods. The first is based on treating the pairwise metric learning problem as a classification problem, where a pair is classified as positive or negative based on the Mahalanobis distance. By observing that the Mahalanobis distance is linear in the entries of M, this leads to a linear classification formulation over pairs. We learn the metric by maximizing the log-likelihood of a logistic discriminant classifier. In (Guillaumin et al., 2010b) instead of M we learn a factorization L, which renders the optimization problem non-convex, but allows to control the number of parameters by learning a rectangular matrix L of size $d \times D$, with $d \ll D$. This is important in case of high-dimensional data, where otherwise we would need *e.g.* a PCA projection to reduce the data dimension, which is sub-optimal since PCA is unsupervised and could discard important data dimensions. A similar metric approach was presented by Mignon and Jurie (Mignon and Jurie, 2012), using a variant of the logistic loss. They showed how to efficiently learn Mahanalobis metrics when the data is represented using kernels. The second method we presented in (Guillaumin et al., 2009b) is a non-parametric method, mKNN, obtained by marginalizing a nearest neighbor classifier. Suppose that we have a training dataset with labeled samples of C classes. We use a k-nearest neighbor classifier to compute the probability that a test sample x_i belongs to class c as $p(y_i = c) = n_{ic}/k$, where n_{ic} is the number
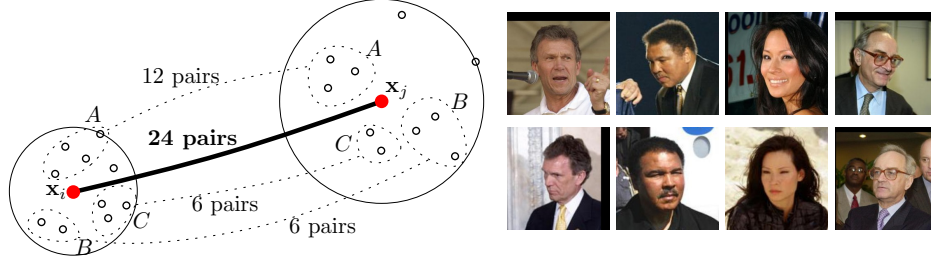
Figure 3.1 – Left: mKNN measures similarity between x_i and x_j by counting the pairs of neighbors with the same class labels. Right: Examples of positive pairs correctly classified using the mKNN classifier with LMNN as a base metric, but wrongly classified using the LMNN metric alone.

of neighbors of x_i of class c. The probability that two samples belong to the same class is then computed by marginalizing over the possible classes that both samples belong to, and given by $p(y_i = y_j) = k^{-2} \sum_c n_{ic} n_{jc}$. Thus, to be similar points do not need to be nearby, as long as they have neighbors of the same classes. In our face verification experiments, this helps in cases where there are extreme pose and expression differences. See Figure 3.1 for an illustration. In (Guillaumin et al., 2010b; Cinbis et al., 2011) we showed how Mahalanobis metrics can also be learned from weakly supervised data, see Section 4.1.

Nearest neighbor prediction models are used in a variety of computer vision problems, including among many others: image location prediction (Hays and Efros, 2008), semantic image segmentation (Tighe and Lazebnik, 2013), and image annotation (Makadia et al., 2010). In nearest neighbor prediction the output is predicted to be one of the outputs associated with each of the neighbors with equal probability. The two hyper-parameters to define are (i) what is the distance measure to define the neighbors, and (ii) how many neighbors to use. In (Guillaumin et al., 2009a) we present a probabilistic nearest neighbor prediction model in which we learn how to weight neighbors, and according to which distance measure to define the neighbors. We will discuss this approach in more detail, and present a selection of experimental results in Section 3.2.

Our model is closely related to the "metric learning by collapsing classes" approach of Globerson & Roweis (Globerson and Roweis, 2006) and the "Large margin nearest neighbor" approach of Weinberger *et al*. (Weinberger et al., 2006). Let us denote the weights over neighbors x_j of a fixed x_i as $\pi_{ij} \propto \exp -d(x_i, x_j)$. When deriving an EM-algorithm for our model, we find an objective function in the M-step that is a KL-divergence between weights π_{ij} and a set of target weights ρ_{ij} computed in the E-step. The ρ_{ij} are large for the x_j nearest to x_i that predict well the output (*e.g.* class label) for x_i. The objective function in (Globerson and Roweis, 2006) is similar but

uses fixed target weights that are uniform for all pairs (i, j) from the same class, and zero for other pairs. The target neighbors in (Weinberger et al., 2006) are defined as the k nearest neighbors of the same class, but they are not updated during learning as the target weights ρ_{ij} in our model.

Many real-life large-scale data collections that can be used to learn image annotation models, such as those constituted by user generated content websites like Flick and Facebook, are open-ended and dynamic: new images are continuously added to existing classes, new classes appear over time, and the semantics of existing classes might evolve too. Most large-scale image annotation and classification techniques rely on efficient linear classification techniques, such as SVM classifiers (Deng et al., 2010; Sánchez and Perronnin, 2011; Lin et al., 2011), and more recently deep convolutional neural networks (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015). To further speed-up the classification, joint dimension reduction and classification techniques have been proposed (Weston et al., 2011), hierarchical classification approaches (Bengio et al., 2011; Gao and Koller, 2011), and data compression techniques (Sánchez and Perronnin, 2011). A drawback of these methods, however, is that when new images become available the classifiers have to be re-trained, or trained from scratch when images of new classes are added.

Distance-based classifiers such as k-nearest neighbors are interesting in this respect, since they enable the addition of new classes and new images to existing classes at negligible computational cost. In (Mensink et al., 2013b) we present a metric learning method for the nearest class mean (NCM) classifier, which avoids the costly neighbor lookup but is a less flexible, linear, classifier as compared to the non-parametric nearest neighbor classifier. We also consider an intermediate approach that represents each class with several centroids, which can represent different sub-classes. A related approach to disambiguate different word senses for keyword based image retrieval was presented in (Lucchi and Weston., 2012). In their work they learn a score function for each query term defined as the maximum over several linear score functions. In our work we learn the centroids in an unsupervised manner, and train a metric used to compute distances to the centroids of all classes. We present this work in more detail in Section 3.3.

Associated publications. We list the most important publications associated with the contributions presented in this chapter here, together with the number of citations they have received.

- (Mensink et al., 2013b) T. Mensink, J. Verbeek, F. Perronnin, G. Csurka. *Distance-based image classification: generalizing to new classes at near-zero cost.* IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (11), pp. 2624–2637, 2013. Citations: 38

- (Mensink et al., 2012) T. Mensink, J. Verbeek, F. Perronnin, G. Csurka. *Metric learning for large scale image classification: generalizing to new classes at near-zero cost.* Proceedings European Conference on Computer Vision, October 2012. Citations: 68

- (Guillaumin et al., 2010b) M. Guillaumin, J. Verbeek, C. Schmid. *Multiple instance metric learning from automatically labeled bags of faces.* Proceedings European Conference on Computer Vision, September 2010. Citations: 78

- (Guillaumin et al., 2009a) M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid. *TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation.* Proceedings IEEE International Conference on Computer Vision, September 2009. Citations: 361

- (Guillaumin et al., 2009b) M. Guillaumin, J. Verbeek, C. Schmid. *Is that you? Metric learning approaches for face identification.* Proceedings IEEE International Conference on Computer Vision, September 2009. Citations: 394

- (Saxena and Verbeek, 2015) S. Saxena, and J. Verbeek. *Coordinated Local Metric Learning.* ICCV ChaLearn Looking at People workshop, December 2015.

3.2   Image annotation with TagProp

In image auto-annotation the goal is to develop methods that can predict for a new image the relevant keywords from an annotation vocabulary (Grangier and Bengio, 2008; Li and Wang, 2008; Liu et al., 2009; Mei et al., 2008). These keyword predictions can be used either to propose tags for an image, or to propose images for a tag or a combination of tags. Non-parametric nearest neighbor like methods have been found to be quite successful for tag prediction (Feng et al., 2004; Jeon et al., 2003; Lavrenko et al., 2003; Makadia et al., 2008; Pan et al., 2004; Zhang et al., 2006; Deng et al., 2010; Weston et al., 2011). This is mainly due to the high 'capacity' of such models: they can adapt flexibly to the patterns in the data as more data is available, without making restrictive linear separability assumptions, as *e.g.* in SVMs. Existing nearest neighbor type methods, however, do not allow for integrated learning of the metric that defines the nearest neighbors in order to maximize the predictive performance of the model. Either a fixed metric (Feng et al., 2004; Zhang et al., 2006) or adhoc combinations of several metrics (Makadia et al., 2008) are used.

In this section we present TagProp, short for "tag propagation", a nearest neighbor image annotation model that predicts tags via weighted predictions from similar training images. The weights are determined either

by the neighbor rank or its distance, and learned via maximum likelihood estimation. This formulation is easily extended to combine several distance functions, *e.g.* based on different features. We also introduce word-specific logistic discriminant models to boost or suppress the tag presence probabilities for very frequent or rare words. This results in a significant increase in the number of words that are predicted for at least one test image.

This work was published in the ICCV'09 paper (Guillaumin et al., 2009a), available here https://hal.inria.fr/inria-00439276/file/GMVS09.pdf.

3.2.1 Weighted nearest neighbor tag prediction

Our goal is to predict the relevance of annotation tags for images. We assume that some visual similarity or distance measures between images are given, abstracting away from their precise definition.

To model image annotations, we use Bernoulli models for each keyword to predict its presence or absence. The dependencies between keywords in the training data are not explicitly modeled, but are implicitly exploited in our model. We use $y_{iw} \in \{-1, +1\}$ to denote the absence/presence of keyword w for image i. The tag presence prediction $p(y_{iw} = +1)$ for image i is a weighted sum over the training images, indexed by j:

$$p(y_{iw} = +1) = \sum_j \pi_{ij} p(y_{iw} = +1|j), \tag{3.1}$$

$$p(y_{iw} = +1|j) = \begin{cases} 1 - \epsilon & \text{if } y_{jw} = +1, \\ \epsilon & \text{otherwise,} \end{cases} \tag{3.2}$$

where π_{ij} denotes the weight of image j for predicting the tags of image i. We require that $\pi_{ij} \geq 0$, and $\sum_j \pi_{ij} = 1$. We use $\epsilon = 10^{-5}$ to avoid zero prediction probabilities. To estimate the parameters that control the weights π_{ij} we maximize the log-likelihood of the predictions of training annotations.

We consider two methods to set the weights for the neighbors: either based on their ranking among the neighbors based on their distance, or directly using the distances instead of the ranks.

Rank-based weights. In the case of rank-based weights over K neighbors we set $\pi_{ij} = \gamma_k$ if j is the k-th nearest neighbor of i. The data log-likelihood is concave in the parameters γ_k, which can be estimated using an EM-algorithm, or a projected-gradient algorithm. The number of parameters equals the neighborhood size K. We refer to this variant as RK, for "rank-based".

This formulation can be easily extended in two ways that are not considered in (Guillaumin et al., 2009a). First, we can exploit multiple similarity measures, *e.g.* based on different features, *i.e.* by defining weights for each combination of rank and similarity measure. Second, the weights can be constrained to be non-increasing with the rank can easily incorporated, since these are linear constraints.

Distance-based weights. Defining the weights directly using distances has the advantage that the weights depend smoothly on the distance, which is important if the distance is to be learned during training. The weights of training images j w.r.t. an image i are in this case defined as:

$$\pi_{ij} = \frac{\exp(-d_\theta(i,j))}{\sum_{j'} \exp(-d_\theta(i,j'))}, \qquad (3.3)$$

where d_θ is a distance metric with parameters θ that we want to optimize. Choices for d_θ include Mahalanobis distances, or positive linear distance combinations of the form $d_\theta(i,j) = \theta^\top d_{ij}$ where d_{ij} is a vector of base distances between image i and j, and the vector θ contains the positive coefficients of the linear distance combination. In our experiments we consider the latter case, in which the number of parameters equals the number of base distances that are combined. When we use a single distance, referred to as the SD variant, θ is a scalar that controls the decay of the weights with distance, and it is the only parameter of the model. When multiple distances are used, the variant is referred to as ML, for "metric learning". We maximize the log-likelihood using a projected gradient algorithm to enforce positivity constraints on the elements of θ. This approach can also be extended to learn Mahalanobis distances, but we did not consider this in our experiments.

Word-specific Logistic Discriminant Models. Weighted nearest neighbor approaches tend to have relatively low recall scores, which is understood as follows. In order to receive a high probability for the presence of a tag, it needs to be present among most neighbors with a significant weight. This is however unlikely to be the case for rare tags, even if some of the neighbors are annotated with the tag, frequent tags are likely to be predicted more strongly.

To overcome this, we introduce word-specific logistic discriminant models that can boost the probability for rare tags and decrease it for very frequent ones. The logistic model uses weighted neighbor predictions by defining

$$p(y_{iw} = +1) = \sigma(\alpha_w x_{iw} + \beta_w), \qquad (3.4)$$

$$x_{iw} = \sum_j \pi_{ij} y_{jw}, \qquad (3.5)$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$ and x_{iw} is the weighted average of annotations for tag w among the neighbors of i, which is equivalent to Eq. (3.1) up to an affine transformation. The word-specific models add two parameters to estimate for each annotation term. We estimate the parameters of the logistic model, and those that determine the neighbor weights in an alternating fashion. We observe rapid convergence, typically after three alternations.

3.2.2 Experimental evaluation

Data sets and experimental setup. We experimented with three publicly available data sets that have been used in previous work, and allow for direct comparison: Corel 5k, ESP Game, and IAPR TC 12. Below we show experimental results for the Corel 5k dataset, and refer to (Guillaumin et al., 2009a) for the results on the other datasets.

We extract different types of features commonly used for image search and categorisation. We use two types of global image descriptors: Gist (Oliva and Torralba, 2001), and color histograms for RGB, LAB, HSV representations. Local features include SIFT (Lowe, 2004) as well as a robust hue descriptor (van de Weijer and Schmid, 2006). By using different color spaces, sampling grids, and possibly including spatial pyramids (Lazebnik et al., 2006), we obtain a total of 15 different image descriptors. For each of these we compute an appropriate distance measure.

We evaluate our models with standard performance measures which evaluate retrieval performance per keyword, and then average over keywords (Carneiro et al., 2007; Feng et al., 2004). Each image is annotated with the 5 most relevant keywords. Then, the mean precision P and recall R over keywords are computed. N+ is used to denote the number of keywords with non-zero recall value. In addition we evaluate precision at different levels of recall as in (Grangier and Bengio, 2008), using mean average precision (mAP) and break-even point precision (BEP).

Experimental results. In our first experiment we compare different variants of TagProp and compare them to the results of the "joint equal contribution" (JEC) model of (Makadia et al., 2008). The latter is essentially a one-nearest-neighbor method that was shown to yield state-of-the-art performance. It determines nearest neighbors using the average of distances computed from different visual features. We re-implemented their method using our own features, referred ot as JEC-15, where we use the average of our 15 normalized base distances to define image similarity.

From the results in Table 3.1 we can make several observations. First, using the tag transfer method proposed in (Makadia et al., 2008) with our own features we obtain similar results. Our rank-based (RK) and distance-based (SD) models that use this fixed distance combination perform com-

	Previsously reported results								TagProp (ours)			
	CRM (Lavrenko et al., 2003)	InfNet(Metzler and Manmatha, 2004)	NPDE (Yavlinsky et al., 2005)	SML (Carneiro et al., 2007)	MBRM (Feng et al., 2004)	TGLM (Liu et al., 2009)	JEC (Makadia et al., 2008)	JEC–15 (ours)	RK	SD	ML	σML
P	16	17	18	23	24	25	27	**28**	28	30	31	**33**
R	19	24	21	29	25	29	32	**33**	32	33	37	**42**
N+	107	112	114	137	122	131	139	**140**	136	136	146	**160**

Table 3.1 – Performance on Corel 5k in terms of P, R, and N+ of our models (using $K = 200$), and those reported in a selection of earlier work. We show results for our variants: RK and SD using the equal distance combination, and ML which integrates metric learning, and σML which further adds the logistic model.

		All- mAP	Single	Multi	Easy	Difficult	All- BEP
PAMIR		26	34	26	43	22	17
TagProp	SD	32	40	31	49	28	24
	σSD	31	41	30	49	27	23
	ML	**36**	43	**35**	53	**32**	**27**
	σML	**36**	**46**	**35**	**55**	**32**	**27**

Table 3.2 – Comparison of TagProp variants (using $K = 200$) and PAMIR in terms of mAP and BEP. The mAP performance is also broken down over single-word and multi-word queries, easy and difficult ones.

parably. When learning the distance combination weights using the ML model significant improvements are obtained, in particular when using the word-specific logistic models (σML). Compared to JEC-15, we obtain marked improvements of 5% in precision, 9% in recall, and count 20 more words with positive recall. This result shows clearly that nearest neighbor type tag prediction can benefit from metric learning.

Above, as most related work, we looked at image retrieval performance for single keywords. Any realistic image retrieval system should, however, support multi-word queries as well. Therefore, we present performance in terms of BEP and mAP on the Corel 5k dataset for both single and multi-word queries. To allow for direct comparison, we follow the setup of (Grangier and Bengio, 2008). Images are considered relevant for a query when they are annotated with all words. The queries are divided into 1,820 'difficult' ones for which there are only one or two relevant images, and 421 'easy' ones with three or more relevant images.

To predict relevance of images for a multi-word query we compute the probability to observe all keywords in the query as the product over the single keyword relevance probabilities according to our model. In Table 3.2 we summarize our results, and compare to those of PAMIR (Grangier and Bengio, 2008) which is a ranking SVM model trained in an online manner. We find that also in this scenario, and for all query types, metric learning improves the results. The word-specific logistic discriminant models are less important in this case, since here we are ranking images for (multi-word) keyword queries, rather than ranking keywords for images. Overall, we gain 10 points in terms of mAP and BEP as compared to PAMIR, which itself was found in (Grangier and Bengio, 2008) to outperform a number of alternative approaches.

3.2.3 Summary

We presented an image annotation model that combines a nearest-neighbor approach with discriminative metric learning. We showed that word-specific logistic discriminant modulation can compensate for varying word fre-

quencies in a data-driven manner. Experimental results show significant improvements over the same model applied to uniformly combined distances. This contrasts with earlier attempts to use metric learning in nearest neighbor image annotation, see *e.g.* (Makadia et al., 2008), that were unsuccessful because the metric was not learned with a method that is coherent with how the metric is used for prediction.

3.3 Metric learning for distance-based classification

In this section we consider large-scale multi-class image classification. We are in particular interested in two distance-based classifiers which enable the addition of new classes and new images to existing classes at negligible computational cost. The k-nearest neighbor (k-NN) classifier is a non-parametric approach that has shown competitive performance for image classification, see Section 3.2 and *e.g.* (Deng et al., 2010). New images (of new classes) are simply added to the dataset, and can be used for classification without further processing. The nearest class mean classifier (NCM) represents classes by their mean feature vector of its elements, see *e.g.* (Webb, 2002). Contrary to the k-NN classifier, which requires (approximate) nearest neighbor look-ups, NCM is an efficient linear classifier. To incorporate new images (of new classes), the relevant class means have to be updated or added to the set of class means.

The success of these methods critically depends on the used distance functions. In our k-NN experiments we use the Large Margin Nearest Neighbor (LMNN) approach (Weinberger et al., 2006) to learn the metric. For the NCM classifier, we propose a novel metric learning algorithm based on multi-class logistic discriminant. Interestingly, in our experiments the NCM classifier is not only more efficient, but also yields better classification accuracy than the k-NN classifier.

The work in this section was first presented in ECCV'12 (Mensink et al., 2012), and an extended version appeared in PAMI (Mensink et al., 2013b). An electronic version of the latter can be found here `https://hal.inria.fr/hal-00817211/file/mensink13pami.pdf`.

3.3.1 Metric learning for the nearest class mean classifiers

We now present our NCM metric learning approach, and an extension to use multiple centroids per class, which transforms the NCM into a more flexible non-linear classifier.

The nearest class mean (NCM) classifier assigns an image to the class c^* with the closest mean: $c^* = \operatorname{argmin}_c d_M(x, \mu_c)$, where $d_M(x, \mu_c)$ is a Mahalanobis distance between an image x and the class mean μ_c. The positive definite matrix M defines the distance metric, and we focus on low-rank

metrics with $M = W^\top W$ and $W \in \mathbb{R}^{d \times D}$, where the rank $d \ll D$ acts as regularizer and reduces the costs of computation and storage. It is easy to verify that this is a linear classifier since, $c^* = \mathrm{argmin}_c\, x^\top w + b$, with $w = -2W^\top W \mu_c$ and $b = \mu_c^\top W^\top W \mu_c$.

We formulate the NCM classifier using a probabilistic model based on multi-class logistic regression and define the probability for a class label c given an feature vector x as:

$$p(c|x) \propto \exp\left(-\tfrac{1}{2} d_W(x, \mu_c) \right). \tag{3.6}$$

This definition may be interpreted as giving the posterior probabilities of a generative model where $p(c)$ is uniform over all classes, and $p(x_i|c) = \mathcal{N}(x_i; \mu_c, \Sigma)$ is a Gaussian with mean μ_c, and a covariance matrix $\Sigma = \left(W^\top W\right)^{-1}$, which is shared across all classes[1].

To learn the projection matrix W, we maximize the log-likelihood of predicting the correct class labels y_i of the training images x_i:

$$\mathcal{L} = \sum_{i=1}^{N} \ln p(y_i | x_i). \tag{3.7}$$

The gradient of this objective function can be written in a simple form as:

$$\nabla_W \mathcal{L} = W \sum_{i=1}^{N} \sum_{c=1}^{C} \alpha_{ic} z_{ic} z_{ic}^\top, \tag{3.8}$$

where $z_{ic} = x_i - \mu_c$, and $\alpha_{ic} = p(c|x_i) - [\![y_i = c]\!]$. The gradient can be interpreted as modifying W to bring the x_i closer to the center of its own class and farther away from the centers of other classes. The scalar weights α_{ic} modulate the terms in the gradient such that most emphasis is on the data points for which the true class is poorly predicted.

Non-linear NCM with multiple centroids per class. To allow for a more expressive model, we can represent each class by a set of centroids, instead of only the class mean. The different centroids per class can be thought of as representing different sub-classes. Let $\{m_{cj}\}_{j=1}^{k}$ denote the set of k centroids for each class c. We define the posterior probability for a centroid m_{cj} as:

$$p(m_{cj}|x) = \frac{1}{Z_x} \exp\left(-\tfrac{1}{2} d_W(x, m_{cj}) \right), \tag{3.9}$$

where $Z_x = \sum_c \sum_j \exp\left(-\tfrac{1}{2} d_W(x, m_{cj}) \right)$ is the normalizer. The posterior probability for class c is then given by:

$$p(c|x) = \sum_{j=1}^{k} p(m_{cj}|x). \tag{3.10}$$

---

[1] Strictly speaking, the covariance matrix is ill defined, since the low-rank matrix $W^\top W$ is non-invertible.

This model corresponds to a generative model where the probability for a feature vector x, to be generated by class c, is given by a Gaussian mixture distribution:

$$p(x|c) = \sum_{j=1}^{k} \pi_{cj} \, \mathcal{N}\left(x_i; m_{cj}, \Sigma\right), \qquad (3.11)$$

with equal mixing weights $\pi_{cj} = 1/k$, and the covariance matrix Σ shared among all sub-classes. We refer to this method as the nearest class multiple centroids (NCMC) classifier.

To learn the projection matrix W, we again maximize the log-likelihood of correct classification. For this model the gradient w.r.t. W is given by:

$$\nabla_W \mathcal{L} = W \sum_{i,c,j} \alpha_{icj} z_{icj} z_{icj}^{\top}, \qquad (3.12)$$

$$z_{icj} = x_i - m_{cj}, \qquad (3.13)$$

$$\alpha_{icj} = p(m_{cj}|x_i) - [\![c = y_i]\!] \frac{p(m_{cj}|x_i)}{\sum_{j'} p(m_{cj'}|x_i)}. \qquad (3.14)$$

The gradient has a similar interpretation as the one derived above for the NCM classifier.

To obtain the centroids of each class, we apply k-means clustering on the features x belonging to that class, using the ℓ_2 distance. The value k offers a transition between NCM ($k = 1$), and a weighted k-NN (k equals the number of images per class), where the weight of each neighbor is defined by the soft-min of its distance, *c.f.* Eq. (3.9). In the limit of large k this model for is similar to TagProp, presented in Section 3.2. The difference in the loss function is that here we consider multi-class image classification, and TagProp (Guillaumin et al., 2009a) was developed for multi-label image annotation.

Large-scale training. For our NCM metric learning approaches, as well as for LMNN, we use SGD training (Bottou, 2010) and sample at each iteration a fixed number of m training images to estimate the gradient. Following (Bai et al., 2010) , we use a fixed learning rate and do not include an explicit regularization term, but rather use the projection dimension d, as well as the number of iterations as an implicit form of regularization.

3.3.2 Experimental evaluation

Datasets, image features, and evaluation measure. In our experiments below we use the dataset of the ImageNet Large Scale Visual Recognition 2010 challenge (ILSVRC'10). To assess performance we report the flat top-5 error rate (lower is better). We extract 4K dimensional Fisher vector (FV)

Projection dim.	32	64	128	256	512	1024	ℓ_2
k-NN	**47.2**	**42.2**	39.7	39.0	39.4	42.4	55.7
NCM	49.1	42.7	39.0	37.4	37.0	**37.0**	68.0
NCMC ($k = 10$)			**35.8**	**34.8**	**34.6**		
WSABIE	51.9	45.1	41.2	39.4	38.7	38.5	

Table 3.3 – Performance of NCM classifiers, as well as k-NN and WSABIE.

(Perronnin et al., 2010b) features computed from local SIFT and color descriptors.

For the k-NN baseline we tune hyper-parameters on the validation set: the number of neighbors, the number of target neighbors in LMNN training, SGD learning rate, and the number of iterations. We also determine the target neighbors of LMNN dynamically in each SGD iteration, which gives an important reduction in the achieved top-5 error rate: *e.g.* from 50.6% to 39.7% when learning a rank 128 metric. For the SVM baseline we follow the one-vs-rest SVM approach of (Perronnin et al., 2012). The top-5 error for the SVM baseline is 38.2%.

Experimental results. In Table 3.3 we show the results obtained with NCM and the related methods for various projection dimensionalities. For both the k-NN and NCM classifiers, using the learned metric outperforms using the ℓ_2 distance by a considerable margin. For k-NN the error rate drops from 55.7% to 39.0%, and for NCM it drops from 68.0% to 37.0%. Perhaps unexpectedly, we observe that our NCM classifier (37.0) outperforms the more flexible k-NN classifier (39.0), as well as the SVM baseline (38.2) when projecting to 256 dimensions or more. Our implementation of WSABIE (Weston et al., 2011) scores slightly worse (38.5), and more importantly it does not generalize to new classes without retraining.

The NCMC classifier that uses multiple centroids per class reduces the error rate further. In Table 3.3 we give results using $k = 10$ centroids per class, which outperforms all other methods (with error 34.6), giving an improvement of 2.4 points over the NCM classifier (37.0), and 3.6 points over SVM classification (38.2).

In (Mensink et al., 2013b) we present experiments with higher dimensional FV features, and comparison to more methods that can generalize to new classes without re-training, including ridge-regression and NCM variants with metrics learned via Fisher linear discriminant analysis and in unsupervised ways. All of these alternatives perform worse than our NCM models evaluated here.

In the second experiment that we highlight here, we use approximately 1M images corresponding to 800 random classes to learn metrics, and evaluate the generalization performance on 200 held-out classes. The error is

	k-NN		NCM			
Projection dim.	128	256	128	256	512	1024
Trained on 800	42.2	42.4	42.5	40.4	39.9	39.6
Trained on all	39.0	38.4	38.6	36.8	36.4	36.5

Table 3.4 – Classification error on images of the 200 classes not used for metric learning, and control setting with metric learning using all classes.

Figure 3.2 – The five nearest classes for two reference classes using the the ℓ_2 distance and a learned metric. See text for details.

evaluated in a 1,000-way classification task, and computed over the 30K images in the test set of the held-out classes. In Table 3.4 we show the performance of NCM and k-NN classifiers, and compare it to the control setting where the metric is trained on all 1,000 classes. For comparison, the one-vs-rest SVM baseline obtains an error of 37.6 on these 200 classes. The results show that both classifiers generalize remarkably well to new classes. For 1024 dimensional projections of the features, the NCM classifier achieves an error of 39.6 over classes not seen during training, as compared to 36.5 when using all classes for training.

Finally, in Figure 3.2, we illustrate the difference between the ℓ_2 and a learned Mahalanobis distance. For two reference classes we show the five nearest classes, based on the distance between class means. We also show the posterior probabilities on the reference class and its five neighbor classes according to Eq. (3.6). The feature vector x is set as the mean of the reference class, *i.e.* a simulated perfectly typical image of this class. We find that the learned metric leads to more visually and semantically related neighbor classes, and much more certain classifications.

3.3.3 Summary

In this section we considered large-scale distance-based image classification, which allow integration of new data and possibly of new classes at a negligible cost. This is not possible with the popular one-vs-rest SVM approach, but is essential when dealing with real-life open-ended datasets. We have introduced a metric learning method for the linear NCM classifier, and presented a non-linear extension based on using multiple centroids per class. We have experimentally validated our models and compared to a state-of-the-art one-vs-rest SVM baseline. Surprisingly we found that the NCM outperforms the more flexible k-NN and that its performance is comparable to a SVM baseline, while projecting the data to as few as 256 dimensions. In (Mensink et al., 2013b) we also present zero-shot learning experiments where we exploit the imagenet class hierarchy to estimate class centroids, and show that NCM provides a unified way to treat classification and retrieval problems.

3.4 Summary and outlook

In this chapter we presented contributions related to metric learning. We highlighted two particular contributions. Our probabilistic weighted nearest neighbor model TagProp offers the advantage that neighbors are not weighted equally, and thus the choice of the number of neighbors is not critical, since far-away neighbors can simply be downweighted so as not to perturb the predictions. The second contribution is a metric learning approach for nearest mean classifier. This classifier predicts class membership based on the distance of a sample to the class mean, using a learned metric. This approach offers improved efficiency w.r.t. a nearest neighbor classifier at test time. In addition, it allows new classes to be added to the model by simply computing the mean of the samples which can be done "on the fly" at negligible cost in practice.

As noted in the introduction of this chapter, similarity and distance measures are of interest for a wide variety of computer vision and other applications. The basic ideas (pairwise and triplet loss functions, based on Euclidean and cosine distances) underlying metric learning can also be applied more or less straightforwardly in the case of deep (convolutional) models, see *e.g.* (Chopra et al., 2005; Schroff et al., 2015). An exiting direction of research, with a significant history in generative modeling see *e.g.* (Hinton et al., 1995; Olshausen and Field, 1997), is to what extent natural image and video structure can be used to learn visual representations with corresponding metrics with supervised learning techniques. For example spatial or temporal proximity (Doersch et al., 2015; Isola et al., 2016; Wang and Gupta, 2015) can be used to define notions of relatedness, which can

be used to learn (deep) visual representations and metrics. While modeling such relations may be of interest by itself, it may also prove useful as an auxiliary task to regularize learning problems with limited supervised training data but very high dimensional parameter space, such as CNNs. Such an approach may be seen as an alternative to the common practice of pre-training on a large supervised dataset, and fine-tuning on the target data (Dosovitskiy et al., 2014; Girshick et al., 2014). In particular, unlike manual supervision, labeling based on spatio-temporal proximity may easily be derived even for non-standard (imaging) sensors, and in much larger quantities even for standard sensors. Moreover, these approaches using unsupervised and supervised data are not mutually exclusive.

Chapter 4

Learning with incomplete supervision

Over the last decade we have witnessed an explosive growth of image and video data available both on-line and off-line. This resulted in the need for tools that automatically analyze the visual content and enrich it with semantically meaningful annotations. Due to the dynamic nature of such archives —new data is added every day— traditional fully supervised machine learning techniques are less suitable. These would require a sufficiently large set of hand-labeled examples of each semantic concept that should be recognized from the low-level visual features. Instead, methods are needed that require less explicit supervision, ideally avoiding any manual labeling of images, and making use of implicit forms of annotation. Examples of implicit annotations are image captions, text associated images on web pages, or scripts, subtitles, or speech transcripts for videos. Such methods offer the hope to leverage the wealth of online visual data to learn visual recognition models. While doing without any manual supervision is a long-term target, in several concrete application areas progress in this direction has been made in recent years.

Contents of this chapter. In Section 4.1 we give a short overview of our contributions in this area in the context of related work in the literature. In Section 4.2 and Section 4.3 we highlight two of our contributions, on structured models for interactive image annotation, and weakly supervised learning for object localization respectively. In Section 4.4 we briefly summarize the contributions from this chapter.

4.1 Contributions and related work

Learning from weaker forms of supervision has become an active and broad line of research, see *e.g.* (Barnard et al., 2003; Fergus et al., 2005; Bekkerman and Jeon, 2007; Li et al., 2007; Papandreou et al., 2015; Pathak et al., 2015; Cinbis et al., 2016b). The crux is to infer the correlations between input data and the missing explicit annotation, based on implicit forms of annotation, *e.g.* from text associated with images, or from subtitles or scripts associated with video (Barnard et al., 2003; Everingham et al., 2006; Satoh et al., 1999; Sivic et al., 2009; Verbeek and Triggs, 2007). The relations that are automatically inferred are necessarily less accurate, than if they were provided by explicit manual annotation efforts. However, weak forms of supervision often comes at a much lower or negligible cost, and therefore typically in much larger volumes. The larger quantity of training data may in practice outweigh the higher quality of fully supervised information.

One of the most used forms of weakly supervised learning has been to exploit the text associated with images on the web. The appearance of web image search engines like Google Images, was rapidly recognized as a way to obtain noisy training examples to learn object recognition models (Berg and Forsyth, 2006; Fergus et al., 2005, 2004). Recently Chatfield *et al.* (Chatfield et al., 2015) have shown that object recognition models can be learned from image search engine results, and applied to retrieve images from collections with millions of images, in a matter of seconds, with low memory footprint, and with high accuracy.

In (Krapac et al., 2010) we developed a model to re-rank web images returned by a visual search engine based on visual and textual consistency, without the need to train a model for every specific query. To enable this we learn a score function over query-relative features, based on training data from a set of diverse queries. Some of these features, for example, indicate whether the query terms appear in various meta-data fields associated with the image, such as in the file name, the web-page title, *etc*. Similarly, visual query-relative features are defined, based on co-occurrence statistics of visual words.

In (Guillaumin et al., 2010a) we developed a semi-supervised method to learn object recognition models from images with user tags, as *e.g.* found on image sharing websites like Flickr. The idea is to learn a strong classifier based on both visual features and tags from a set of labeled images. This strong classifier is then used to assess unlabeled images that also come with tags. Finally, a visual-only classifier is learned from both the labeled and unlabeled images. This improves the performance of the final classifier as compared to using only the labeled images, since the tag information is leveraged at training time to indentify additional unlabeled examples.

A related line of research considers interactive learning and classificatin methods to maximally exploit a small amount of manual annotation ef-

fort. Active learning methods, see *e.g.* (Settles, 2009; Vijayanarasimhan and Grauman, 2011), interleave model updates with requesting users to annotate images which, given the current model, are likely to be the most effective to improve the model. Others have considered to use user interaction to improve automatic predictions for difficult fine-grained classification problems, such as recognizing bird species (Branson et al., 2010). The observation is that while recognizing the bird species might be difficult, it is easy for users to give input on visual attributes (*e.g.*, the color of the beak). The user-provided attributes are then used to narrow-down the possible taget classes.

In our own work (Mensink et al., 2011, 2013a) we considered a similar problem of interactive image annotation, where the goal is to optimally predict all relevant image labels from a minimum amount of user input. By using a structured model over the image labels the user input of one label can be propagated to better predict other labels, and to identify the most useful labels for further user input. We present this work in more detail in Section 4.2.

Another example of weakly supervised learning is the learning face recognition models from image captions (Berg et al., 2004), or subtitle and script information (Everingham et al., 2009). In the case of still images, face detections are associated with names that are detected in image captions. Similarly in video, detected faces are tracked over time, and the face tracks are associated with speaker names indicated in the script. The script can be temporally aligned with the video by relying on subtitles which, unlike scripts, have a precise temporal anchoring in the video. In both cases the problem is formulated as a matching problem between a set of tentative names for a set of detected faces. The main difficulty is to overcome the appearance variability of the same person due to changes in viewpoint, lighting, and expression.

In (Guillaumin et al., 2008; Mensink and Verbeek, 2008; Guillaumin et al., 2012) we developed matching methods based on similarity graphs (maximizing the weight of edges among nodes assigned to the same person), and using classifiers (interleaving training the classifiers and assigning faces to the mostly likely classes). To obtain an effective measure of face similarity despite the chalenges metioned above, we used our logistic discriminant metic learning approach (Guillaumin et al., 2009b) to learn a Mahalanobis metric. To learn the metric, however, requires labeled face images. In (Guillaumin et al., 2010b) we consider learning such metrics directly from weakly supervised captioned images using a multiple instance learning approach. In (Cinbis et al., 2011) we use temporal constraints to learn a metric from unsupervised face tracks obtained from video. To form positive and negative face pairs for metric learning, we use the fact that all faces in a track belong to the same person, and that face tracks that occur simultaneously in time depict different people.

For object localization, weakly supervised learning from image-wide labels that indicate the presence of instances of a category in images has recently been intensively studied as a way to remove the need for bounding box annotations, see *e.g.* (Bagon et al., 2010; Chum and Zisserman, 2007; Crandall and Huttenlocher, 2006; Deselaers et al., 2012; Pandey and Lazebnik, 2011; Prest et al., 2012; Russakovsky et al., 2012; Shi et al., 2013; Siva et al., 2012; Siva and Xiang, 2011; Song et al., 2014a,b; Bilen et al., 2014; Wang et al., 2014a). While earlier work was based on datasets where the viewpoint changes were controlled, *e.g.* the training images consisted only of sideviews of cars (Deselaers et al., 2010), more recent work has moved to more challenging datasets which are not viewpoint constrained (Siva and Xiang, 2011). Most of the existing work takes a multiple instance learning (MIL) approach, where learning the detector is interleaved with inferring the most likely object location in each positive training image.

In (Cinbis et al., 2014) we proposed a novel MIL learning approach which avoids some of the poor local optima that are recovered by standard MIL training. This is particularly important when using high-dimensional image representations such as the Fisher vector. Moreover, we also propose a window refinement procedure, which encourages the object hypotheses to better align with full object outligns rather than with discriminative parts. In Section 4.3 we will present this work in more detail.

Weakly supervised object localization has also been studied in the video domain. For example, Prest *et al*. (Prest et al., 2012) propose to learn object recognition models from weakly supervised videos. They cluster long-term optical flow based trajectories to segment the video in several parts of coherent motion (Brox and Malik, 2010), which are used as candidate locations for object localization. In (Oneata et al., 2014a) we proposed a different method to generate object localization candidates in video based on hierarchical supervoxel video segmentations.

Closely related to weakly supervised object localization are the tasks of co-segmentation (Joulin et al., 2010) and co-localization (Joulin et al., 2014), where only a set of positive images that contain the object class of interest is used to jointly localize the object instances across the images in terms of a segmentation or bounding-box localization. Recent work has reported encouraging results in an even more challenging scenario (Cho et al., 2015), where the training set consists of images that contain instances of multiple object categories, without supervised information which category is present in which image.

Another area where weakly supervised learning is attractive is semantic image segmentation (Shotton et al., 2006). Here the goal is to label each image pixel with a category label. Clearly, obtaining training images with complete pixel-level labelings is a time comsuming process. To alleviate the labeling effort, we have developed sematic segmentation models that can be trained either from images where only a subset of the pixels is labeled
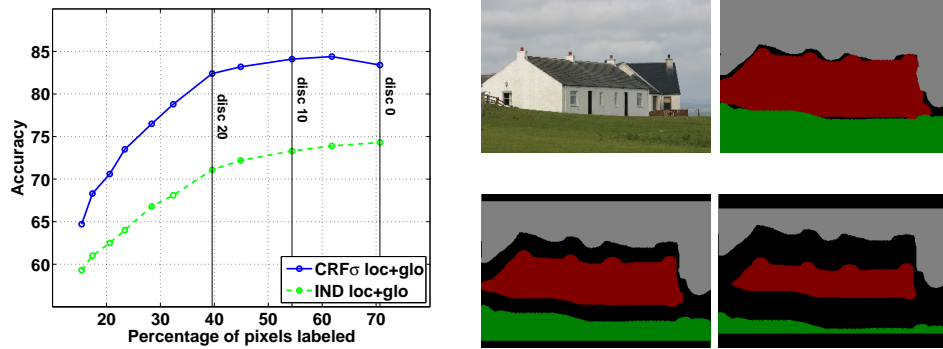
Figure 4.1 – Per-pixel recognition accuracy when learning from increasingly eroded label maps (left). Example image with its original label map, and erosions thereof with disk of size 10 and 20 (right). The missing labels are inferred using loopy belief propagation during training. The CRF model gives significantly better accuracy than the "IND" model that predicts labels independently.

(notably without any labeled pixels at the category boundaries) (Verbeek and Triggs, 2008), and when using only image-wide labels that indicate which categories are present in the image (Verbeek and Triggs, 2007). We used generative and discriminative random field models that use unary potentials to guide the local category recognition, and pairwise potentials to ensure spatial contiguity of the labeling. In addition, in (Verbeek and Triggs, 2007), we used a global potential in the form of a sparse Dirichle prior, that encourages the labeling to be sparse in the sense that in each image only a small number of all posible categories are used in the labeling. See Figure 4.1 for an illustration of results we obtained when learning from incomplete label maps in (Verbeek and Triggs, 2008). Very recently (Papandreou et al., 2015; Pathak et al., 2015) significant progress has been made on this problem, by learning CNN models for semantic segmentation from image-level labels. The main contribution in these works is the use of constraints that force that at least a certain fraction of the pixels in an image is labeled with each of the image-wide labels.

Associated publications. We list our most important publications associated with the contributions presented in this chapter here, together with the number of citations they have received.

- (Cinbis et al., 2016b) G. Cinbis, J. Verbeek, C. Schmid. *Weakly supervised object localization with multi-fold multiple instance learning.* IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear, 2016. Citations: 2

- (Mensink et al., 2013a) T. Mensink, J. Verbeek, G. Csurka. *Tree-structured CRF models for interactive image labeling.* IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2), pp. 476–489, 2013. Citations: 20

- (Guillaumin et al., 2012) M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid. *Face recognition from caption-based supervision.* International Journal of Computer Vision, 96(1), pp. 64–82, January 2012. Citations: 54

- (Cinbis et al., 2014) G. Cinbis, J. Verbeek, C. Schmid. *Multi-fold MIL Training for Weakly Supervised Object Localization.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2014. Citations: 30

- (Cinbis et al., 2011) G. Cinbis, J. Verbeek, C. Schmid. *Unsupervised metric learning for face identification in TV video.* Proceedings IEEE International Conference on Computer Vision, November 2011. Citations: 69

- (Mensink et al., 2011) T. Mensink, J. Verbeek, G. Csurka. *Learning structured prediction models for interactive image labeling.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2011. Citations: 26

- (Guillaumin et al., 2010a) M. Guillaumin, J. Verbeek, C. Schmid. *Multimodal semi-supervised learning for image classication.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2010. Citations: 241

- (Krapac et al., 2010) J. Krapac, M. Allan, J. Verbeek, F. Jurie. *Improving web image search results using query-relative classifiers.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2010. Citations: 86

- (Guillaumin et al., 2008) M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid. *Automatic face naming with caption-based supervision.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, June 2008. Citations: 77

- (Mensink and Verbeek, 2008) T. Mensink, and J. Verbeek. *Improving people search using query expansions: How friends help to find people.* Proceedings European Conference on Computer Vision, pp. 86–99, October 2008. Citations: 38

- (Verbeek and Triggs, 2008) J. Verbeek and B. Triggs. *Scene segmentation with CRFs learned from partially labeled images.* Advances

in Neural Information Processing Systems 20, pp. 1553–1560, January 2008. Citations: 121

- (Verbeek and Triggs, 2007) J. Verbeek and B. Triggs. *Region classification with Markov field aspect models.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, June 2007. Citations: 241

4.2 Interactive annotation using label dependencies

Most existing systems address image annotation either fully manually (*e.g.* stock photo sites as Getty images, http://www.gettyimages.com) or fully automatically (where image labels are automatically predicted without any user interaction). In the latter case most commonly used are either classifiers, *e.g.* (Zhang et al., 2007), ranking models *e.g.* (Grangier and Bengio, 2008), or nearest neighbor predictors (Guillaumin et al., 2009a). The vast majority of these methods do not explicitly model dependencies among the image labels.

In this section we consider structured models that explicitly take into account the dependencies among image labels. We follow a semi-automatic labeling scenario, where test images are annotated based on partial user input for a few image labels. This is, for example, useful when indexing images for stock photography, where a high annotation quality is mandatory, yet fully manually indexing is very expensive and suffers from low throughput. Label dependencies can be leveraged in two ways. First, to transfer the user input for one image label to more accurate predictions on other image labels. Second, to identify those image labels for user input that are most informative on the remaining image labels.

The material here appeared initially at CVPR'11 (Mensink et al., 2011) and later in extended form in PAMI (Mensink et al., 2013a). An reprint of the latter is available at https://hal.inria.fr/hal-00688143/file/MVC2012pami.pdf

4.2.1 Tree-structured image annotation models

Our goal is to model dependencies between image labels, but which allows for tractable inference. To this end, we define a tree-structured conditional random field model, where each node represents a label from the annotation vocabulary, and edges between nodes represent interaction terms between the labels. Let $\boldsymbol{y} = (y_1, \ldots, y_L)^\top$ denote a vector of the L binary label variables, *i.e.* $y_i \in \{0, 1\}$. We define the probability for a specific configuration $\boldsymbol{y}$ given the image $\boldsymbol{x}$:

$$p(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\big(-E(\boldsymbol{y}, \boldsymbol{x})\big), \tag{4.1}$$

where $E(\boldsymbol{y}, \boldsymbol{x})$ is an energy function scoring the compatibility between an image $\boldsymbol{x}$ and a label vector $\boldsymbol{y}$.

The label tree is defined by a set of edges $\mathcal{E} = \{e_1, \ldots, e_{L-1}\}$, where $e_l = (i, j)$ indicates an edge between y_i and y_j. For a given tree structure the energy for a configuration of labels $\boldsymbol{y}$ for an image $\boldsymbol{x}$ is given by:

$$E(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{L} \psi_i(y_i, \boldsymbol{x}) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(y_i, y_j). \quad (4.2)$$

For the unary terms we use generalized linear functions:

$$\psi_i(y_i = l, \boldsymbol{x}) = \boldsymbol{\phi}_i(\boldsymbol{x})^\top \boldsymbol{w}_i^l, \quad (4.3)$$

where $\boldsymbol{\phi}_i(\boldsymbol{x})$ is a feature vector for the image which may depend on the label index i, and $\boldsymbol{w}_i^l$ is the weight vector for state $l \in \{0, 1\}$. In particular, we set $\boldsymbol{\phi}_i(\boldsymbol{x}_n) = [s_i(\boldsymbol{x}_n), 1]^\top$, where $s_i(\boldsymbol{x})$ is the score of an SVM classifier for label y_i that is obtained using a method reminiscent of cross-validation. We also experimented with setting $\phi_i(x)$ to the FV features used by the SVMs, but found this to be less effective. See (Mensink et al., 2013a) for details.

The pairwise potentials, defined by a scalar parameter for each joint state of the corresponding nodes, are independent of the image input:

$$\psi_{ij}(y_i = s, y_j = t) = v_{ij}^{st}. \quad (4.4)$$

Given the tree structure, we learn the parameters of the unary and pairwise potentials by the maximum likelihood criterion. As the energy function is linear in the parameters, the log-likelihood function is concave and the parameters can be optimized using gradient-based methods. Computing the gradient requires evaluation of the marginal distributions on single variables and pairs of variables connected by edges in the tree. These can be efficiently obtained in time linear in the number of image labels using belief propagation due to the tree structure (Pearl, 1982).

4.2.2 Obtaining the structure of the model

The interactions between the labels are defined by the structure of the tree. Finding the optimal tree structure for conditional models is generally intractable (Bradley and Guestrin, 2010), therefore we have to resort to approximate methods to determine the structure of the tree. We use the optimal tree structure for a generative model instead, which can be found using the Chow-Liu algorithm (Chow and Liu, 1968) as the maximum spanning tree in a fully connected graph over the label variables with edge weights given by the mutual information between the label variables. As an alternative to the Chow-Liu algorithm, we experimented with a greedy
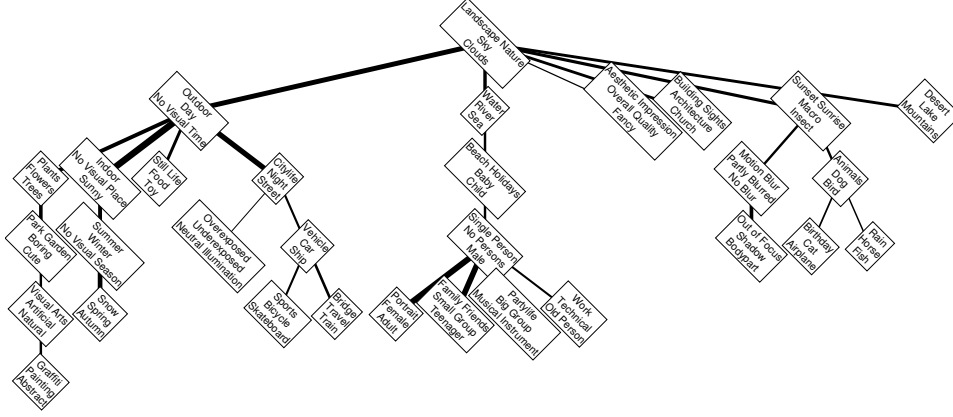
Figure 4.2 – An example tree over compound nodes with $k = 3$ labels on the 93 labels of the ImageCLEF data set. The edge width is proportional to the mutual information between the linked nodes. The root of the tree has been chosen as the vertex with highest degree.

maximum-likelihood method to learn the tree structure, but did not find it to give significantly better structures (Mensink et al., 2013a).

To allow for richer dependencies, we define trees over label groups instead of individual labels. To obtain the label groups, we perform agglomerative clustering based on mutual information, fixing in advance a maximum group size k. We determine a tree structure on the compound nodes as before using the Chow-Liu algorithm. In Figure 4.2 we show a tree with group size $k = 3$, which shows that semantically related concepts are often grouped together.

In order to be less dependent on a particular choice for the size of the label groups, we combine tree-structured models over label groups of different sizes. The models are combined in a mixture, where each tree defines a mixture component which gives a joint distribution over the labels. We train the trees independently, and mix their predictions using uniform mixing weights.

4.2.3 Label elicitation for image annotation

In the semi-automatic image annotation scenario, a user is asked to state for one or more labels if they are relevant to the image. The question is: which are the most useful labels to be presented to the user? We propose a label selection strategy whose aim is to minimize the uncertainty of the remaining labels given the test image. This strategy resembles those used for query selection in active learning (Settles, 2009).

The uncertainty of the remaining labels given the value of y_i can be quantified by the conditional entropy. Since the value of y_i is not known

user input, we instead compute the expected conditional entropy

$$H(\boldsymbol{y}_{\backslash i}|y_i, \boldsymbol{x}) = \sum_l p(y_i = l|\boldsymbol{x})H(\boldsymbol{y}_{\backslash i}|y_i = l, \boldsymbol{x}), \qquad (4.5)$$

where $\boldsymbol{y}_{\backslash i}$ denotes all label variables except y_i. Using the fact that $H(\boldsymbol{y}|\boldsymbol{x})$ does not depend on the selected variable y_i, and given the basic identity of conditional entropy, see *e.g.* (Bishop, 2006), we have

$$H(\boldsymbol{y}|\boldsymbol{x}) = H(y_i|\boldsymbol{x}) + H(\boldsymbol{y}_{\backslash i}|y_i, \boldsymbol{x}). \qquad (4.6)$$

We conclude that minimizing Eq. (4.5) for y_i is equivalent to maximizing $H(y_i|\boldsymbol{x})$ over i. Hence, we select the label y_{i*} with $i^* = \arg\max_i H(y_i|\boldsymbol{x})$ to be set by the user. When using mixtures of trees, a similar analysis can be used to quantify the conditional entropy in terms of label uncertainties.

In order to select multiple labels to be set by the user, we proceed sequentially by first asking the user to set only one label. We then repeat the procedure while conditioning on the input already provided by the user.

4.2.4 Experimental evaluation

Datasets and evaluation measures. We experimented with three datasets, for more details on these datasets and comparison of the results to the literature we refer to (Mensink et al., 2013a). In the *ImageCLEF'10* dataset (Nowak and Huiskes, 2010) the images are labeled with 93 diverse concepts, see Figure 4.2. The *SUN'09* dataset (Choi et al., 2010) contains 107 labels (107), with around 5 labels per image on average, this is significantly more than in the PASCAL VOC 2007 data set which has only 20 labels and over 50% of the images having only a single label. The *Animals with Attributes (AwA)* (Lampert et al., 2009b) data set contains images of 50 animal classes, and a definition of each class in terms of 85 attributes. In the experiments reported here, we predict the attribute annotations for this dataset.

We measure the performance of the methods using: (i) *MAP*, a retrieval performance measure, which is the mean average precision (AP) over all keywords, where AP is computed over the ranked images for a given keyword, and (ii) *iMAP*, the mean AP over all images, where AP is computed over the ranked labels for an image.

Experimental results. We compare an independent label prediction model, tree-structured models with different node sizes, and mixture of such trees in Figure 4.3. In the fully automatic label prediction setting (first row), we observe that the MAP/iMAP performance of the structured prediction models is about $1 - 1.5\%$ higher than of the independent model. The performance differences between the models with different group sizes k can be interpreted as a trade-off between model capacity and overfitting. For all data sets the mixture-of-trees performs the best.
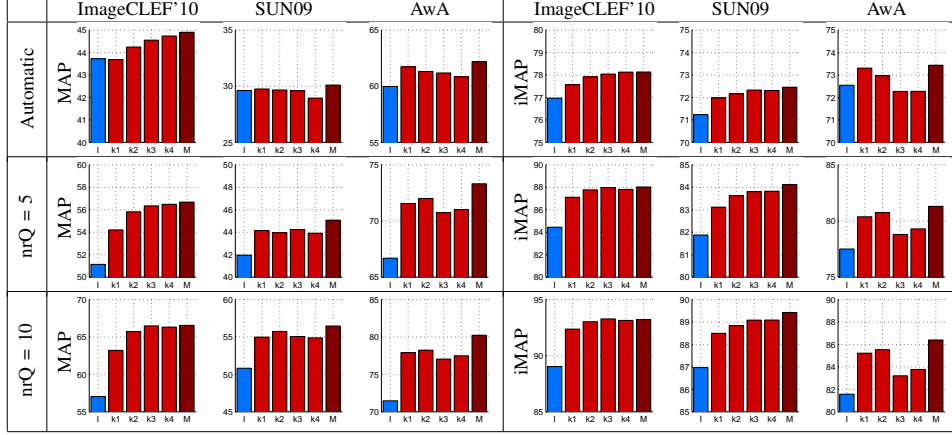
Figure 4.3 – Performance for fully automated prediction (first row), and an interactive setting with 5 and 10 questions (second and third row). For each setting and dataset, we compare results of the independent model (I, blue), the trees with group sizes k from 1 to 4 (k1–k4, light-red), and the mixture-of-trees (M, dark-red).

In the interactive image annotation scenario the system iteratively selects labels to be set by the user (set to the ground value in our experiments). For the independent model, the entropy-based selection procedure is also used, which results in setting the most uncertain labels. The annotation results obtained after setting 5 respectively 10 labels are shown in the second and third rows of Figure 4.3. Note the different vertical scales across the different rows. As expected, in this setting the structured models benefit more from the user input, since they propagate the information provided by the user to update the predictions on the remaining labels, and also avoids asking input for multiple highly correlated labels. The mixture-of-trees again performs optimal, or close to optimal, in all cases.

To assess the proposed label elicitation method we compare its performances to using a random strategy, we do so using the independent model and the mixture-of-trees model For the random strategy we report the average performance over ten experiments. The results in Figure 4.4 show that with either elicitation mechanism, the structured model outperforms the independent model. Furthermore, for both models the entropy-based label elicitation mechanism is more effective than random selection.

4.2.5 Summary

In this section we presented tree-structured models to capture dependencies among image labels. We explored (i) different strategies to learn the unary potentials (pre-trained SVM classifiers and joint learning with the
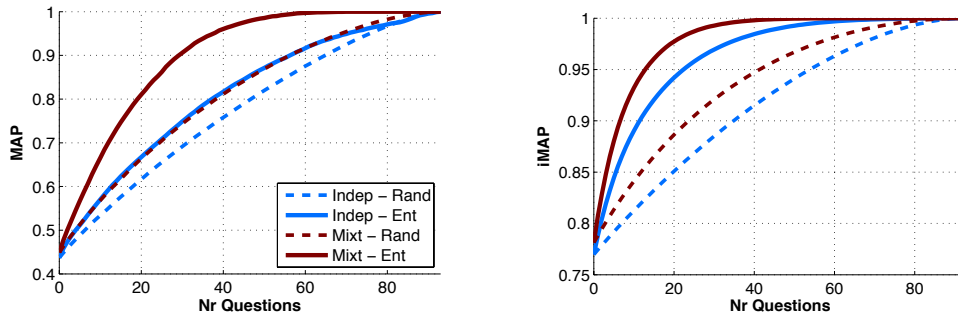
Figure 4.4 – Comparison of the random and entropy-based label selection for the independent and the structured mixture-of-trees model using the ImageCLEF'10 dataset.

pairwise potentials), (ii) various graphical structures (trees, trees over label groups, and mixtures of trees), and (iii) methods to obtain these structures (using mutual information and based on maximum likelihood). We find that best performance is obtained using a mixture-of-trees with different label group sizes, where the unary potentials are given by pre-trained SVM classifiers. During training, the SVM scores are obtained in a cross-validation manner, to ensure that the quality of the SVM scores is representative of that of test images. The proposed models offer a moderate improvement over independent baseline models in a fully automatic setting. Their main strength lies in improved predictions in an interactive image labeling setting.

4.3 Weakly supervised learning for object localization

For object detection, weakly supervised learning from image-wide labels that indicate the presence of instances of a category in images has recently been intensively studied as a way to remove the need for bounding box annotations, see *e.g.* (Bagon et al., 2010; Chum and Zisserman, 2007; Crandall and Huttenlocher, 2006; Deselaers et al., 2012; Pandey and Lazebnik, 2011; Prest et al., 2012; Russakovsky et al., 2012; Shi et al., 2013; Siva et al., 2012; Siva and Xiang, 2011; Song et al., 2014a,b; Bilen et al., 2014; Wang et al., 2014a). In this section, we present a method based on multiple instance learning that interleaves training of the detector with re-localization of object instances on the positive training images. Following recent state-of-the-art work in fully supervised detection (Cinbis et al., 2013; Girshick et al., 2014; van de Sande et al., 2014), we represent tentative detection windows using high-dimensional Fisher vectors (140K dims.) (Sánchez et al., 2013) and convolutional neural network features (140K dims.) (Krizhevsky et al., 2012). When used in an MIL framework, the high-dimensionality of the

window features makes MIL quickly convergence to poor local optima after initialization. Our main contribution is a multi-fold training procedure for MIL, which avoids this rapid convergence to poor local optima. In addition, we propose a window refinement method that improves the weakly supervised localization accuracy by incorporating a category-independent objectness measure.

Part of this material was presented at CVPR'14 (Cinbis et al., 2014), an extended version of the paper will appear in PAMI (Cinbis et al., 2016b). The latter is available at https://hal.inria.fr/hal-01123482/file/paper_final.pdf.

4.3.1 Multi-fold training for weakly supervised localization

The majority of related work treats WSL for object detection as a multiple instance learning (MIL) (Dietterich et al., 1997) problem. Each image is considered as a "bag" of examples given by tentative object windows. Positive images are assumed to contain at least one positive object instance window, while negative images only contain negative windows. The object detector is then obtained by alternating detector training, and using the detector to select the most likely object instances in positive images.

In many MIL problems, *e.g.* such as those for weakly supervised face recognition (Berg et al., 2004; Everingham et al., 2009), the number of examples per bag is limited to a few dozen at most. In contrast, there is a vast number of examples per bag in the case of object detector training since the number of possible object bounding boxes is quadratic in the number of image pixels. Object detection proposal methods, *e.g.* (Alexe et al., 2010; Gu et al., 2012; Uijlings et al., 2013; Zitnick and Dollár, 2014), can be used to make MIL approaches to WSL for object localization manageable, and make it possible to use powerful and computationally expensive object models. In our work we use the selective search method of Uijlings *et al*. (Uijlings et al., 2013), to generate a limited set of around 1,500 candidate windows per image. Jointly selecting the objects across the retained windows across thousands of images, however, is still a challenging problem since the number of choices is exponential in the number of images.

Note that in the MIL approach described above, the detector used for re-localization in positive images is trained using positive samples that are extracted from the very same images. Therefore, there is a bias towards re-localizing on the same windows; in particular when high capacity classifiers are used which are likely to separate the detector's training data. For example, when a nearest neighbor classifier is used the re-localization will be degenerate and not move away from its initialization, since the same window will be found as its nearest neighbor. The same phenomenon occurs when using powerful and high-dimensional image representations to train linear classifiers. We illustrate this in the left panel of Figure 4.5, which
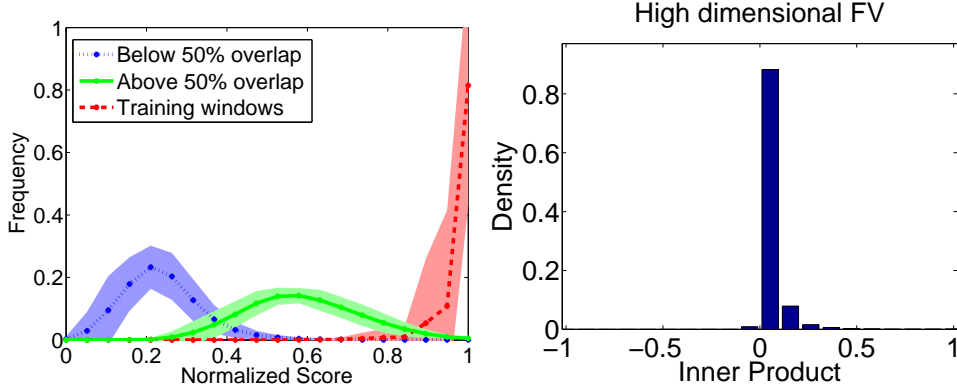
Figure 4.5 – Left: distribution of the window scores in positive training images during MIL training. Red: windows used for training. Green: other windows that overlap with them by more than 50%. Blue: windows that overlap less than 50%. Each curve is obtained by averaging all per-class score distributions. The surrounding regions show the standard deviation. Right: distribution of inner products between Fisher vectors of pairs of windows, where each pair is sampled from within a single image.

shows the distribution of the window scores in a typical MIL iteration on VOC 2007 using Fisher vectors. We observe that the windows used in SVM training score significantly higher than the other ones, including those with a significant spatial overlap with the most recent training windows. As a result, MIL typically results in degenerate re-localization.

This problem is related to the dimensionality of the window descriptors. We illustrate this in the right panel of Figure 4.5, where we show the distribution of inner products between the descriptors of window pairs within the same image. Almost all window descriptors are near orthogonal for the 140K dimensional FVs. Recall that the weight vector of a linear SVM classifier can be written as a linear combination of training samples, $\boldsymbol{w} = \sum_i \alpha_i \boldsymbol{x}_i$, and the SVM score of a test sample is given by a linear combination of inner products with training vectors. Therefore, the training windows are likely to score significantly higher than the other windows in positive images in the high-dimensional case, resulting in degenerate re-localization behavior.

Note that increasing regularization weight in SVM training does not remedy this problem. The ℓ_2 regularization term with weight λ restricts the linear combination weights such that $|\alpha_i| \leq 1/\lambda$. Therefore, although we can reduce the influence of individual training samples via regularization, the resulting classifier remains biased towards the training windows since the classifier is a linear combination of the window descriptors.

To address this issue—without sacrificing the descriptor dimensional-

ity, which would limit its descriptive power—we propose to train the detector using a multi-fold procedure, reminiscent of cross-validation, within the MIL iterations. We divide the positive training images into K disjoint folds, and re-localize the images in each fold using a detector trained using windows from positive images in the other folds. In this manner the re-localization detectors never use training windows from the images to which they are applied. Once re-localization is performed in all positive training images, we train another detector using all selected windows. This detector is used for hard-negative mining on negative training images, and returned as the final detector.

The number of folds used in our multi-fold MIL training procedure should be set to strike a trade-off between two competing factors. On the one hand, using more folds increases the number of training samples per fold, and is therefore likely to improve re-localization performance. On the other hand, using more folds increases the computational cost.

4.3.2   Window refinement

An inherent difficulty for weakly supervised object localization is that WSL labels only permit to determine the most repeatable and discriminative patterns for each class. Therefore, even though the windows found by WSL are likely to overlap with target object instances, they might not align with the full object outline. We propose a window refinement method to update the localizations obtained by multi-fold training. The final detector is trained based on these refinements.

To explicitly take into account object boundaries, we use the edge-driven objectness measure of Zitnick and Dollar (Zitnick and Dollár, 2014). The main idea of this approach is to score a given window based on the number of contours that are fully contained inside the window, with an increased weight on near-boundary edge pixels. Thus, windows that tightly enclose long contours are scored highly, whereas those with predominantly straddling contours are penalized. Additionally, in order to reduce the effect of slight misalignments, the coordinates of a given window are updated using a greedy local search procedure that aims to increase the objectness score.

In (Zitnick and Dollár, 2014), the objectness measure is used for generating object proposals. We instead use the edge-driven objectness measure to improve WSL outputs. For this purpose, we combine the objectness measure with the detection scores given by multi-fold MIL. More specifically, we first utilize the local search procedure in order to update and score the refined candidate detection windows based on the objectness measure, without updating the detection scores. To make the detection and objectness scores comparable, we scale both scores to the range $[0, 1]$ for all windows in the positive training images. We then average both scores, and select the top detection in each image with respect to this combined score.
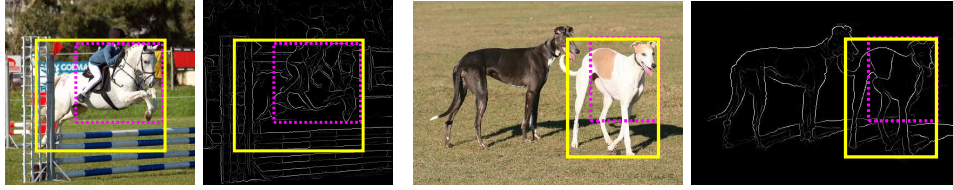
Figure 4.6 – Illustration of window refinement. Dashed pink boxes show the localization before refinement, and the solid yellow boxes show the result after refinement. The right-most image in each pair shows the edge map used to compute the objectness measure.

In order to avoid selecting the windows irrelevant for the target class, but with a high objectness score, we restrict the search space to the top-N windows per image in terms of the detection score. While we use $N = 10$ in all our experiments, we have empirically observed that the refinement method significantly improves the localization results for N ranging from 1 to 50. The improvement is comparable for $N \geq 5$.

In Figure 4.6, we show example images for the classes *horse* and *dog* together with the corresponding edge maps. In these images, the dashed (pink) boxes show the output of multi-fold MIL training and the solid (yellow) boxes show the outputs of the window refinement procedure. Even though the initial windows are located on the object instances, they are evaluated as incorrect due to the low overlap ratios with the ground-truth ones. The edge maps show that many contours, *i.e.* most object contours, straddle the initial window boundaries. In contrast, the refined windows have higher percentages of fully contained contours, *i.e.* the contours relevant for the objects.

4.3.3 Experimental evaluation

For our experiments we used the PASCAL VOC 2007 dataset (Everingham et al., 2010). We use linear SVM classifiers, and set the weight of the regularization term and the class weighting to fixed values based on preliminary experiments. We perform two hard-negative mining steps (Felzenszwalb et al., 2010) after each re-localization phase.

Following (Deselaers et al., 2012), we assess performance using two measures. First, we evaluate the fraction of positive *training images* in which we obtain correct localization (CorLoc). Second, we measure the final object detection performance on the *test images* using the standard protocol (Everingham et al., 2010): average precision (AP) per class, summarized by the mean AP (mAP) across all 20 classes. For both measures, a window is considered correct if it has an intersection-over-union ratio of at least 50% with a ground-truth object.

	Corloc			mAP		
	FV	CNN	FV+CNN	FV	CNN	FV+CNN
Standard MIL	29.7	41.2	34.4	15.5	24.3	22.0
Multi-fold MIL	38.8	45.0	47.3	22.4	25.9	27.4
+Refinement	46.1	54.2	52.0	23.3	28.6	30.2

Table 4.1 – Comparison of standard and multi-fold MIL training, and the effect of window refinement. Performance both in CorLoc on the positive training images (left), and in mAP on the test images. Results are averaged over the 20 object categories in PASCAL VOC'07.

In Table 4.1 we give a brief summary of the results of the extensive set of experiments we conducted, more details can be found in (Cinbis et al., 2016b). We report the CorLoc and mAP values across all classes for both the FV and CNN features, as well as their combination. In all settings, and according to both measures, both the multi-fold training procedure and the window refinement bring significant improvements to the performance of the detectors. The improvement due to multi-fold training is more pronounced when using the 140K dimensional FV representation. The CNN descriptors are only 4K dimensional, and are therefore to a lesser degree affected by the near-orthogonality of window descriptors observed in Figure 4.5.

Our results are comparable to the current state of the art. For example Bilen and Vedaldi (Bilen and Vedaldi, 2016) report 30.6% mAP and 51% CorLoc using a two-stream CNN approach based on the same detection proposal windows, which finetune the CNN weights.

4.3.4 Summary

We presented a multi-fold multiple instance learning approach for weakly supervised object detection. It improves localization performance by separating the image sets for re-localization and model training. We also presented a window refinement method, which improves the localization accuracy by using an edge-driven objectness prior.

We have evaluated our approach and compared it to state-of-the-art methods using the VOC 2007 dataset. Our results show that multi-fold MIL effectively handles high-dimensional descriptors, which allows us to obtain results that are competitive with the state of the art by combining FV and CNN features.

A detailed analysis of our results shows that, in terms of test set detection performance, multi-fold MIL attains 68% of the MIL performance upper-bound, which we measure by selecting one correct training example from each positive image, for the combined FV and CNN features.

4.4 Summary and outlook

In this chapter we presented an overview of our contributions related to learning visual models from incomplete supervision, and highlighted two of them. In the first we model semantic image label dependencies, which allows us to leverage user provided information on part of the labels to better predict the remaining unknown ones. The model also allows to infer which labels are most informative when given by the user. Experimental result demonstrate the effectiveness of this model for interactive image labeling. The second contribution is a multi-fold multiple-instance learning framework, which we apply to learning object category localization models from weakly supervised data. In this case the training data only indicates if an object category is present in an image, but now where.

While fully supervised methods to learn visual recognition models deliver the best performance in general, they come with the important drawback of requiring large and carefully annotated datasets. Collecting such datasets in practice is often time consuming, expensive, and not-trivial to setup. As an example consider collecting annotations for semantic video segmentation, where full supervision requires labeling each pixel in each frame with a corresponding category label. Learning from incomplete forms of supervision is an important topic of research in computer vision and machine learning in general, which can alleviate the costs of collecting supervised datasets. The advent of deep visual recognition models only underlines the importance of this issue, due to their large number of parameters. Weakly supervised learning is most often expressed as learning parameters of latent variable models, where the latent variables correspond to the missing supervision. Learning is then done with algorithms such as Expectation-Maximization (Dempster et al., 1977), or simple variants such as multiple instance learning (Dietterich et al., 1997). Latent variable models beyond tree-structured factor graphs require approximate inference techniques, see *e.g.* (Verbeek and Triggs, 2008), and the effect of the precise inference method on the learned model are relatively poorly understood (Kulesza and Pereira, 2008). Recent work (Zheng et al., 2015; Schwing and Urtasun, 2015) interprets variational mean-field inference as a recurrent neural network through which error signals can be back-propagated. This ensures that the model parameters are learned directly to predict well when combined with the chosen inference method. Generalization of this principle is an interesting line for future work that could address the following questions. How to re-formulate more powerful approximate inference methods, such as generalized loopy belief propagation (Yedidia et al., 2002), or expectation propagation (Minka, 2001), as recurrent networks? How to incorporate higher-order potential functions in such approaches, beyond very specific restrictive classes (Arnab et al., 2015).

Chapter 5

Conclusion and perspectives

In this concluding chapter we summarize the contributions described in the previous chapters in Section 5.1, and identify several long-term research directions in Section 5.2.

5.1   Summary of contributions

Below we briefly review the previous chapters, and discuss related directions for future research.

Fisher vector representations.   In Chapter 2 we discussed our contributions around the Fisher vector (FV) image representation in the context of related work. These include the derivation of the Fisher information matrix w.r.t. the mixing weights in (Sánchez et al., 2013), modeling the layout of local descriptors with a FV that represents the distribution of spatial coordinates of each visual word (Krapac et al., 2011), and using approximate segmentation masks to weight the contribution of local descriptors in the FV for object localization (Cinbis et al., 2013). In (Cinbis et al., 2012, 2016a) we presented models for local image descriptors that avoid the i.i.d. assumption that underlies the BoV and FV representations. These models naturally lead to discounting effects and consequent performance improvements, which are comparable to those obtained using power normalization. Using our models we can interpret power-normalization as an approximate manner to account for mutual dependencies of local descriptors. In (Oneata et al., 2014b) we presented approximations to the power and ℓ_2 normalizations of the FV. Using these approximations linear score functions of the normalized FV can be computed efficiently using integral images, since the interaction of the weight vector with local descriptors is additive per visual word. Experimental results show that a speed-up of more than an order of magnitude is obtained, while having only a limited impact on localization performance.

The Fisher kernel framework has shown to be one of the most effective methods to encode the distribution of local features in images and videos (Chatfield et al., 2011; Oneata et al., 2013). Recently there has been a major focus in computer vision on convolutional neural network (CNN) approaches following the success of such models at the 2012 ImageNet challenge (Krizhevsky et al., 2012). Recent work also explored hybrid approaches that combine aspects of local feature pooling and (convolutional) neural networks (Cimpoi et al., 2015; Perronnin and Larlus, 2015; Arandjelović et al., 2015). In particular, using a FV to aggregate local convolutional filters learned with a CNN, was shown in (Cimpoi et al., 2015) to improve over using higher-level CNN represenatations for transfer learning problems. We believe that developing Fisher kernels for generative models that capture more structural aspects is an interesting direction of future research. Recent examples in this direction include (Sun and Nevatia, 2013; Nagel et al., 2015), which uses a FV representation for video event recognition based on a hidden Markov models and Student-t distributions, and (Sánchez and Redolfi, 2015) which derives general exponential family FV representations *e.g.* to model positive definite matrices, or binary data.

Metric learning approaches. We presented our contributions related to metric learning in Chapter 3. In (Guillaumin et al., 2009b) we presented LDML, a logistic discriminant Mahanalobis metric learning approach, and a non-parametric marginalized nearest neighbor approach. We extended LDML in (Guillaumin et al., 2010b) to learn low-rank Mahalanobis metrics, and to use it in combination with kernel functions. In (Guillaumin et al., 2009a) we presented a nearest neighbor image annotation model, where instead of using equal weights for a fixed number of neighbors, we use a weighted combination of predictions made by neighboring images. We set the weights either based on the neighbor rank, or based on a learned combination of several distance metrics between images. In (Mensink et al., 2013b) we presented a Mahalanobis metric learning approach for the nearest class mean (NCM) classifier. Unlike the nearest neighbor (NN) classifier, this is an efficient linear classifier. We also considered a non-linear extension where each class is represented with several centroids that can represent sub-classes. In our experiments we found NCM to outperform NN classification, while at the same time also being computationally more efficient.

While most work on metric learning considers a supervised setting, it is also possible to learn metrics from unsupervised data. For example, in (Cinbis et al., 2011) we used face tracks in videos in combination with simple temporal constraints to derive training examples for LDML metric learning. Similarly, co-occurrence statistic have been used to learn vectorial word representations from unsupervised text corpora. For example,

the skip-gram model (Mikolov et al., 2013a,b) learns a word embedding so that words that frequently occur nearby in text are also co-located in the learned embedding. It is an interesting direction of future research to explore similar approaches to learn metrics for visual representations. For example, we can learn a metric and corresponding data represenation so that video frames that appear nearby in the same video tend to be close according to the learned metric, and that frames of different videos tend to be far apart. We expect to be able to learn high-level semantic representations in this manner, since even if the objects depicted in nearby video frames might be completely disjoint, we still expect the visual content to be semantically related if they are sampled relatively nearby in time from the same video. Recent examples of work along these lines include (Doersch et al., 2015; Isola et al., 2016; Wang and Gupta, 2015; Dosovitskiy et al., 2014; Zou et al., 2012). The motivation underlying these works is that natural visual data exhibits many structural regularities, which may be exploited to learn representations or to regularize supervised learning. This is a particularly relevant line of work in the current era of powerful (convolutional) neural networks, which have extremely large numbers of parameters, and which are non-trivial to learn and regularize.

Learning with incomplete supervision. In Chapter 4 we presented our contributions related to learning from incomplete supervision. These include image re-ranking models that generalize to new queries (Krapac et al., 2010), semi-supervised image classification models that leverage user provided keywords for training (Guillaumin et al., 2010a), approaches to associate names and faces in captioned news images and in videos (Guillaumin et al., 2008; Mensink and Verbeek, 2008; Guillaumin et al., 2010b; Cinbis et al., 2011; Guillaumin et al., 2012), and semantic image segmentation models that can be learned from incomplete supervision (Verbeek and Triggs, 2007, 2008). We developed tree-structured models over labels for interactive image annotation (Mensink et al., 2011, 2013a) exploiting keyword dependencies to gather more informative user input and improve predictions. Finally, we developed a multi-fold multiple instance learning approach for weakly supervised object localization (Cinbis et al., 2014, 2016b), which avoids poor local optima during learning and consequently improves the localization performance.

In ongoing research we work on learning semantic video segmentation models from weak supervision, including separately segmenting individual category instances. Recent advances in object localization and semantic segmentation have revealed a number of effective techniques, which are yet to be combined in a larger overall model. These include pooling operators over variable-sized areas (Ren et al., 2015; He et al., 2014), fully connected CRFs (Krähenbühl and Koltun, 2011) and convolutional and decon-

volutional computation of unary potentials (Long et al., 2015; Ronneberger et al., 2015; Noh et al., 2015), non-trivial data-dependent and trainable pairwise potentials (Lin et al., 2016), recurrent networks for approximate variational inference intregrated in the training process (Schwing and Urtasun, 2015; Zheng et al., 2015), and the use of (linearly) constrained variational inference for weakly supervised learning (Pathak et al., 2015). Object localization models, possibly learned from image-wide labels as in (Cinbis et al., 2014), can be used to define strong prior distributions for semantic segmentation, *e.g.* as in (Ladický et al., 2010). Moreover, the strong temporal correlation patterns in the label maps in semantic video segmentation suggests the use of recurrent models to exploit this structure.

5.2 Long-term research directions

We now conclude with several more general long-term research directions.

Learning higher-order structured prediction models. Many problems in computer vision involve joint prediction of many response variables. Examples include, but are not limited to, semantic segmentation, optical flow estimation, depth estimation, image de-noising, super resolution, colorization, pose estimation, *etc*. These structured prediction tasks are typically solved using (conditional) Markov random fields, which includes unary terms for each label variable, and pairwise terms to ensure structural regularity of the output predictions.

Deep networks have been used for such tasks (Long et al., 2015) to define unary and pairwise terms (Lin et al., 2016). Deep networks allow complex functions to be learned between a label variable and a large part of all input variables, if not all. Moreover, recently (Zheng et al., 2015; Schwing and Urtasun, 2015) it has been shown that variational mean-field inference in Markov random fields can be expressed as a special recurrent neural networks in the case of fully connected pair-wise energy functions. This allows the training of the unary and pairwise potentials to be done in a way that is coherent with the MRF structure and the approximate inference method.

Higher-order potentials, that model interactions of more than two label variables at a time, have been proven effective in the past for structured prediction tasks, see *e.g.* (Kohli et al., 2009). Efficient inference, however, is only possible for a very specific classes of higher-order potentials, see *e.g.* (Vineet et al., 2014; Ramalingam et al., 2008). An exiting direction for future work is to consider how larger classes of trainable higher-order potentials can be used by generalizing the techniques developed in (Zheng et al., 2015; Schwing and Urtasun, 2015) for pairwise structured models. The work of Pinheiro and Collobert on recurrent convolutional networks (Pinheiro and Collobert, 2014) is also highly relevant in this area. An al-

ternative route to enforce higher-order regularity in the predictions might be to used adversarial networks (Goodfellow et al., 2014) that are trained in combination with the primary prediction model. The adversarial network is trained to discriminate ground-truth samples and samples from the primary model. The primary model is trained such that the adversarial network can not discriminate samples from the primary model from samples of the ground-truth. The adversarial network may be used to enforce higher-order consistency, even if higher-order potentials are not used in the primary model. The development of models that exhibit high-order regularities which are trainable in a data-driven manner are likely to have an significant impact across a wide variety of multivariate and dense prediction vision problems.

Learning from minimal supervision. An important bottleneck limiting performance of visual recognition systems in practical applications is the reliance on supervised training dataset. Generally, supervision is expensive and time consuming to collect. There are at least three different paths to make up for a lack of supervised training data.

The first is to learn models that to go beyond recognizing (*i.e.* classifying, localizing, segmenting, *etc.*) a manually specified finite list of (object) categories. Approaches in this direction include semantic word-image embedding models such as DeViSE (Frome et al., 2013), and image-caption encoder-decoder models (Kiros et al., 2015). Such models can in principle be learned from large non-curated datasets which contain images with (loosely) associated textual descriptions (general web images, wikipedia, user generated content, *etc.*), see *e.g.* (Chen et al., 2013b). This approach, combined with word-embedding techniques (Mikolov et al., 2013b) and "on-the-fly" model learning from web image-search engines (Chatfield et al., 2015), allows to learn bi-directional image-text mappings that can be used for example for free-text visual search in large image and video datasets, without requiring any manually curated supervised training datasets.

Second, for certain critical visual recognition tasks that require high-level of accuracy (*e.g.* advanced driver assistance systems, or defense related applications), manually collected supervised training datasets will be required to ensure sufficient accuracy. In such cases the question is how we can make the most out of the (limited) available training data. An idea that has proven extremely effective is to use auxiliary tasks to pre-train or initialize the recognition model, see *e.g.* (Girshick et al., 2014). Most often pre-training is based on large supervised training datasets; with ImageNet (Deng et al., 2009) being by far the most used dataset for this purpose. Large unsupervised datasets may also be used for this purpose, by defining auxiliary tasks based on spatial or temporal structure (Doersch et al., 2015; Wang and Gupta, 2015; Isola et al., 2016). Most work takes a rather ad-hoc

approach of taking a pre-trained model, and adapting it to the task at hand. In a more principled manner, we can learn by jointly minimizing the loss of the (new) target task and a loss for the (earlier) auxiliary task(s). Pushing this idea further, a "life-long" learning scheme is interesting in which we train a single large model for an increasing number of tasks. Treating the "old" tasks a pre-training or regularization for the new tasks.

Finally, a third approach is to rely on contextual cues. These can either in the form of spatial inter-object context, see *e.g.* (Rabinovich et al., 2007; Choi et al., 2010), or between objects and physical scene properties such as scene geometry estimates, see *e.g.* (A. Geiger and Urtasun, 2011; Hoiem et al., 2008). Another form of context is to use complex data adaptive non-parametric priors on the parameters of discriminative recognition models, see *e.g.* (Salakhutdinov et al., 2012). Such priors can infer hierarchical groupings of object categories, so that training data is shared to some extent between related classes.

These three paragraphs may be summarized as follows. (i) For some problems abundantly available and loosely annotated training data may be enough to learn satisfying models, *e.g.* for text-based image search. (ii) In cases where this is not sufficient, auxiliary tasks may be used for pre-training, or multi-task learning can be used as a regularization principle to make up for lack of supervised training data. (iii) Contextual information of various forms can provide stronger structuring information. Future research on combining these different approaches may lead to important advances in learning visual recognition models from very little training data, which may have significant impact for practical applications.

Architecture learning and adaptation. Current state of the art high-level semantic scene understanding models are dominated by (convolutional) neural network approaches. These models are very powerful due to their strong capacity to model complex data distributions, which results from a hierarchical structure with millions configurable parameters that can be automatically tuned based on (supervised) training data (Montufar et al., 2014). Beyond the challenges to efficiently estimate such models from limited training data, an even bigger challenge is posed by the model selection problem. That is: how to determine the best, or a "good", architecture for such models? This includes: the number and ordering of pooling and convolutional layers, filter sizes, number of channels, type of pooling operations, type of non-linearities, *etc*. This problem is extremely hard, since the space of possible network architectures is discrete and combinatorially large. Optimizing over this space is an important challenge for future research. Work in this direction includes using sparsity inducing regularizers to sparsify the connectivity pattern (Kulkarni et al., 2015), and using sparse hierarchical priors over the network structure in a Bayesian learning frame-

work (Adams et al., 2010).

In the context of extremely large datasets, such as those used for learning from weakly supervised sources discussed above, model selection might not be the right problem to consider. Instead of searching for the single ultimate model architecture, it will be important to progressively adapt the model architecture and capacity during learning. That is: having seen little data it might be useful to limit the degrees of freedom of the model. As the learning algorithm sees more data the limited capacity will saturate, and more capacity should be allocated. This suggests that studying a dynamic variant of the model selection problem is perhaps more important.

The model selection problem is highly challenging, but progress is likely to have big impact across many computer vision problems and beyond.

Bibliography

The zettabyte era: Trends and analysis. White Paper, 2015. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.pdf.

C. Wojek A. Geiger and R. Urtasun. Joint 3d estimation of objects and scene layout. In *NIPS*, 2011.

R. Adams, H. Wallach, and Z. Ghahramani. Learning the structure of deep sparse graphical models. In *AISTATS*, 2010.

B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010.

B. Alexe, T. Deselares, and V. Ferrari. Measuring the objectness of image windows. *PAMI*, 34(11):2189–2202, 2012.

R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.

R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. Arxiv preprint, 2015.

A. Arnab, S. Jayasumana, S. Zheng, and P. Torr. Higher order potentials in end-to-end trainable conditional random fields. 2015. URL http://arxiv.org/abs/1511.08119.

S. Bagon, O. Brostovski, M. Galun, and M. Irani. Detecting and sketching the common. In *CVPR*, 2010.

B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Weinberger. Learning to rank with (a lot of) word features. *Information Retrieval*, 13(3):291–314, 2010.

K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003.

R. Bekkerman and J. Jeon. Multi-modal clustering for multimedia collections. In *CVPR*, 2007.

A. Bellet, A. Habrard, and M. Sebban. A Survey on Metric Learning for Feature Vectors and Structured Data. *ArXiv e-prints*, 1306.6709, 2013.

S. Bengio, J. Weston, and D. Grangier. Label embedding trees for large multi-class tasks. In *NIPS*, 2011.

T. Berg and D. Forsyth. Animals on the web. In *CVPR*, 2006.

T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *CVPR*, 2004.

H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with posterior regularization. In *BMVC*, 2014.

C. Bishop. *Pattern recognition and machine learning.* Spinger-Verlag, 2006.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, 2010.

J. Bradley and C. Guestrin. Learning tree conditional random fields. In *ICML*, 2010.

S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.

T. Brox and J. Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010.

G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *PAMI*, 29(3):394–410, 2007.

K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.

K. Chatfield, R. Arandjelović, O. Parkhi, and A. Zisserman. On-the-fly learning for visual search of large-scale image and video datasets. *International Journal of Multimedia Information Retrieval*, 2015.

Q. Chen, Z. Song, R. Feris, A. Datta, L. Cao, Z. Huang, and S. Yan. Efficient maximum appearance search for large-scale object detection. In *CVPR*, 2013a.

X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge from web data. In *ICCV*, 2013b.

M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild. In *CVPR*, 2015.

M. Choi, J. Lim, A. Torralba, and A. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, 2005.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Information Theory*, 14(3):462–467, 1968.

O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, 2007.

M. Cimpoi, S. Maji, and A. Vedaldi. Deep filter banks for texture recognition and segmentation. In *CVPR*, 2015.

R. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in TV video. In *ICCV*, 2011.

R. Cinbis, J. Verbeek, and C. Schmid. Image categorization using Fisher kernels of non-iid image models. In *CVPR*, 2012.

R. Cinbis, J. Verbeek, and C. Schmid. Segmentation driven object detection with Fisher vectors. In *ICCV*, 2013.

R. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014.

R. Cinbis, J. Verbeek, and C. Schmid. Approximate Fisher kernels of non-iid image models for image categorization. *PAMI*, 2016a.

R. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *PAMI*, 2016b. to appear.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, 1995.

D. Crandall and D. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *ECCV*, 2006.

G. Csurka and F. Perronnin. An efficient approach to semantic segmentation. *IJCV*, 95(2):198–212, 2011.

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Int. Workshop on Stat. Learning in Computer Vision*, 2004.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. doi: 10.1109/CVPR.2005.177. URL http://hal.inria.fr/inria-00548512.

J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

J. Deng, A. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.

T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *ECCV*, 2010.

T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):257–293, 2012.

T. Dietterich, R. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89 (1-2):31–71, 1997.

C. Doersch, A. Gupta, and A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

A. Dosovitskiy, J. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *NIPS*, 2014.

O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *ICCV*, 2009.

M. Everingham, J. Sivic, and A. Zisserman. 'Hello! My name is... Buffy' - automatic naming of characters in TV video. In *BMVC*, 2006.

M. Everingham, J. Sivic, and A. Zisserman. Taking the bite out of automatic naming of characters in TV video. *Image and Vision Computing*, 27(5): 545–559, 2009.

M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, June 2010.

P. Felzenszwalb, R. Grishick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010.

S. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

R. Fergus, P. Perona, and A. Zisserman. A visual category filter for Google images. In *ECCV*, 2004.

R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *ICCV*, 2005.

B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, 2015.

A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. DeViSE: A deep visual-semantic embedding model. In *NIPS*, 2013.

T. Gao and D. Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *ICCV*, 2011.

R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

A. Globerson and S. Roweis. Metric learning by collapsing classes. In *NIPS*, 2006.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

D. Grangier and S. Bengio. A discriminative kernel-based model to rank images from text queries. *PAMI*, 30(8):1371–1384, 2008.

K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. Draw: A recurrent neural network for image generation view publication. In *icml*, 2015.

C. Gu, P. Arbeláez, Y. Lin, K. Yu, and Malik. Multi-component models for object detection. In *ECCV*, 2012.

M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Automatic face naming with caption-based supervision. In *CVPR*, 2008.

M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009a.

M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. In *ICCV*, 2009b.

M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, 2010a.

M. Guillaumin, J. Verbeek, and C. Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *ECCV*, 2010b.

M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Face recognition from caption-based supervision. *IJCV*, 96(1):64–82, 2012.

J. Hays and A. Efros. im2gps: estimating geographic information from a single image. In *CVPR*, 2008.

K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014.

G. Hinton, P. Dayan, B. Frey, and R. Neal. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268:1158–1161, 1995.

D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 80:3–15, 2008.

P. Isola, D. Zoran, D. Krishnan, and E. Adelson. Learning visual groups from co-occurrences in space and time. In *ICLR*, 2016.

T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, 1999.

H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, 2009.

H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *PAMI*, 34(9):1704–1716, 2012.

J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *ACM SIGIR*, 2003.

Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14, 2014.

M. Jordan, editor. *Learning in Graphical Models*. Kluwer, 1998.

M. Jordan, Z. Ghahramani, T. Jaakola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image cosegmentation. In *CVPR*, 2010.

A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with Frank-Wolfe algorithm. In *ECCV*, 2014.

F. Khan, R. Anwer, J. van de Weijer, A. Bagdanov, M. Vanrell, and A. Lopez. Color attributes for object detection. In *CVPR*, 2012.

R. Kiros, R. Salakhutdinov, and R. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015. to appear.

Takumi Kobayashi. Dirichlet-based histogram feature transform for image classification. In *CVPR*, 2014.

P. Kohli, L. Ladický, and P. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(3):302–324, 2009.

M. Köstinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

J. Krapac, M. Allan, J. Verbeek, and F. Jurie. Improving web-image search results using query-relative classifiers. In *CVPR*, 2010.

J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with Fisher vectors for image categorization. In *ICCV*, 2011.

A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

A. Kulesza and F. Pereira. Structured learning with approximate inference. In *NIPS*, 2008.

B. Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012.

P. Kulkarni, J. Zepeda, F. Jurie, P. Pérez, and L. Chevallier. Learning the structure of deep architectures using l1 regularization. In *BMVC*, 2015.

L. Ladický, P. Sturgess, K. Alahari, C. Russell, and P. Torr. What, where & how many? combining object detectors and crfs. In *ECCV*, 2010.

C. Lampert, M. Blaschko, and T. Hofmann. Efficient subwindow search: a branch and bound framework for object localization. *PAMI*, 31(12): 2129–2142, 2009a.

C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009b.

I. Laptev and P. Pérez. Retrieving actions in movies. In *ICCV*, 2007.

V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. In *NIPS*, 1989.

T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001.

J. Li and J. Wang. Real-time computerized annotation of pictures. *PAMI*, 30 (6):985–1002, 2008.

L.-J. Li, G. Wang, and L. Fei-Fei. OPTIMOL: Automatic object picture collection via incremental model learning. In *CVPR*, 2007.

Z. Li, E. Gavves, K. van de Sande, C. Snoek, and A. Smeulders. Codemaps, segment classify and search objects locally. In *ICCV*, 2013.

G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016.

Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale image classification: Fast feature extraction and SVM training. In *CVPR*, 2011.

J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recognition*, 42(2):218–228, 2009.

J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

A. Lucchi and J. Weston. Joint image and word sense discrimination for image retrieval. In *ECCV*, 2012.

A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, 2008. URL http://www.cis.upenn.edu/~makadia/annotation/.

A. Makadia, V. Pavlovic, and S. Kumar. Baselines for image annotation. *IJCV*, 90(1):88–105, 2010.

T. Mei, Y. Wang, X.S. Hua, S. Gong, and S. Li. Coherent image annotation by learning semantic distance. In *CVPR*, 2008.

T. Mensink and J. Verbeek. Improving people search using query expansions: How friends help to find people. In *ECCV*, 2008.

T. Mensink, J. Verbeek, and G. Csurka. Learning structured prediction models for interactive image labeling. In *CVPR*, 2011.

T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *ECCV*, 2012.

T. Mensink, J. Verbeek, and G. Csurka. Tree-structured CRF models for interactive image labeling. *PAMI*, 35(2):476–489, 2013a.

T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *PAMI*, 35(11):2624–2637, 2013b.

D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *CIVR*, 2004.

A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013a.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013b.

T. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, Massachusetts, USA, 2001.

G. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *NIPS*, 2014.

M. Nagel, T. Mensink, and C. Snoek. Event Fisher vectors: Robust encoding visual diversity of visual streams. In *BMVC*, 2015.

H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

S. Nowak and M. Huiskes. New strategies for image annotation: Overview of the photo annotation task at ImageCLEF 2010. In *Working Notes of CLEF*, 2010.

Y. Ohta, T. Kanade, and T. Sakai. An analysis system for scenes containing objects with substructures. In *ICPR*, 1978.

A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311 – 3325, 1997.

D. Oneata. *Robust and efficient models for action recognition and localization*. PhD thesis, Université de Grenoble, 2015.

D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with Fisher vectors on a compact feature set. In *ICCV*, 2013.

D. Oneata, J. Revaud, J. Verbeek, and C. Schmid. Spatio-temporal object detection proposals. In *ECCV*, 2014a.

D. Oneata, J. Verbeek, and C. Schmid. Efficient action localization with approximately normalized Fisher vectors. In *CVPR*, 2014b. submitted.

P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. Smeaton, and G. Quénot. TRECVID 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID*, 2012.

J. Pan, H. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *ACM SIGKDD*, 2004.

M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

G. Papandreou, L.-C. Chen, K. Murphy, and A. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015.

D. Pathak, P. Krahenbühl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.

J. Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the Second National Conference on Artificial Intelligence*, 1982.

X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked Fisher vectors. In *ECCV*, 2014.

A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic. A hybrid generative/discriminative classification framework based on free energy terms. In *ICCV*, 2009a.

A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic. Free energy score space. In *NIPS*, 2009b.

F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

F. Perronnin and D. Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In *CVPR*, 2015.

F. Perronnin, J. Sánchez, and Y. Liu. Large-scale image categorization with explicit data embedding. In *CVPR*, 2010a.

F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010b.

F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid. Towards good practice in large-scale learning for image classification. In *CVPR*, 2012.

P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *ICML*, 2014.

A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012.

A. Rabinovich, A. Vedaldi, C. Galleguillos, and E. Wiewiora S. Belongie. Objects in context. In *ICCV*, 2007.

S. Ramalingam, P. Kohli, K. Alahari, and P. Torr. Exact inference in multi-label CRFs with higher order cliques. *CVPR*, 2008.

S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. URL http://arxiv.org/abs/1506.01497.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *ECCV*, 2012.

R. Salakhutdinov, J. Tenenbaum, and A. Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In *ICML Unsupervised and Transfer Learning workshop*, 2012.

J. Sánchez and F. Perronnin. High-dimensional signature compression for large-scale image classification. In *CVPR*, 2011.

J. Sánchez and J. Redolfi. Exponential family Fisher vector for image classification. *Pattern Recognition Letters*, 59:26 – 32, 2015.

J. Sánchez, F. Perronnin, and T. de Campos. Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognition Letters*, 33(16): 2216–2223, 2012.

J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

S. Satoh, Y. Nakamura, and T. Kanade. Name-it: Naming and detecting faces in news videos. *IEEE MultiMedia*, 6(1):22–35, 1999.

S. Saxena and J. Verbeek. Coordinated local metric learning. In *ICCV ChaLearn Looking at People workshop*, 2015.

C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19(5):530–534, 1997.

F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

A. Schwing and R. Urtasun. Fully connected deep structured networks. Arxiv preprint, 2015.

B. Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison, 2009.

Z. Shi, T. Hospedales, and T. Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *ICCV*, 2013.

J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, pages 1–15, 2006.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

K. Simonyan, O. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, 2013.

P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *ECCV*, 2012.

Parthipan Siva and Tao Xiang. Weakly supervised object detector learning with model drift detection. In *ICCV*, 2011.

J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *ICCV*, 2003.

J. Sivic, M. Everingham, and A. Zisserman. "Who are you?": Learning person specific classifiers from video. In *CVPR*, 2009.

H. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *ICML*, 2014a.

H. Song, Y. Lee, S. Jegelka, and T. Darrell. Weakly-supervised discovery of visual pattern configurations. In *NIPS*, 2014b.

C. Sun and R. Nevatia. ACTIVE: activity concept transitions in video event classification. In *ICCV*, 2013.

J. Tighe and S. Lazebnik. Superparsing - scalable nonparametric image parsing with superpixels. *IJCV*, 101(2):329–349, 2013.

J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013.

K. van de Sande, C. Snoek, and A. Smeulders. Fisher and VLAD with FLAIR. In *CVPR*, 2014.

J. van de Weijer and C. Schmid. Coloring local feature extraction. In *ECCV*, 2006.

J. van Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *PAMI*, 32(7):1271–1283, 2010.

A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010.

J. Verbeek and B. Triggs. Region classification with Markov field aspect models. In *CVPR*, 2007.

J. Verbeek and B. Triggs. Scene segmentation with CRFs learned from partially labeled images. In *NIPS*, 2008.

S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR*, 2011.

V. Vineet, J. Warrell, and P. Torr. Filter-based mean-field inference for random fields with higher-order terms and product label-spaces. *IJCV*, 2014.

P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 57(2):137–154, 2004.

C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*, 2014a.

H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013. URL http://hal.inria.fr/hal-00803241.

H. Wang, D. Oneata, J. Verbeek, and C. Schmid. A robust and efficient video representation for action recognition. *IJCV*, 2015.

J. Wang, K. Sun, F. Sha, S. Marchand-Maillet, and A. Kalousis. Two-stage metric learning. In *ICML*, 2014b.

X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015.

A. Webb. *Statistical pattern recognition*. Wiley, New-York, NY, USA, 2002.

K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009.

K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

J. Weston, S. Bengio, and N. Usunier. WSABIE: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.

J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.

A. Yavlinsky, E. Schofield, and S. Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *CIVR*, 2005. URL www.edschofield.com/publications/yavlinsky05automated.pdf.

J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. Technical report, Mitsubishi Electric Research Laboratories, 2002.

J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *CVPR*, 2009.

H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, pages 2126–2136, 2006.

J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *IJCV*, 73(2):213–238, 2007.

S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.

C. Zitnick and P. Dollár. Edge boxes: locating object proposals from edges. In *ECCV*, 2014.

W. Zou, S. Zhu, A. Ng, and K. Yu. Deep learning of invariant features via simulated fixations in video. In *NIPS*, 2012.

Appendix A

Curriculum vitae

INRIA Rhône-Alpes
Tel. +33 4 76 61 52 33
655 Avenue de l'Europe, 38330 Montbonnot, France

Email: Jakob.Verbeek@inria.fr
Webpage: http://thoth.inrialpes.fr/∼verbeek
Citizenship: Dutch, Date of birth: December 21, 1975

Curriculum Vitae – Jakob Verbeek

Academic Background

2004   • Doctorate Computer Science (best thesis award), Informatics Institute, University of Amsterdam. Advisors: Prof. Dr. Ir. F. Groen, Dr. Ir. B. Kröse, and Dr. N. Vlassis. Thesis: *Mixture models for clustering and dimension reduction.*

2000   • Master of Science in Logic (with honours), Institute for Language, Logic, and Computation, University of Amsterdam. Advisor: Prof. Dr. M. van Lambalgen. Thesis: *An information theoretic approach to finding word groups for text classification.*

1998   • Master of Science in Artificial Intelligence (with honours), Dutch National Research Institute for Mathematics and Computer Science & University of Amsterdam. Advisors: Prof. Dr. P. Vitányi, Dr. P. Grünwald, and Dr. R. de Wolf. Thesis: *Overfitting using the minimum description length principle.*

Awards

2011   • Outstanding Reviewer Award, IEEE Conference on Computer Vision and Pattern Recognition.

2009   • Outstanding Reviewer Award, IEEE Conference on Computer Vision and Pattern Recognition.

2006   • Biannual E.S. Gelsema Award of the Dutch Society for Pattern Recognition and Image Processing for best PhD thesis and associated international journal publications.

2000   • Regional winner of yearly best MSc thesis award Dutch Society for Computer Science.

Employment

since 2007   • Researcher (CR1), INRIA Rhône-Alpes, Grenoble.

2005-2007   • Postdoc, INRIA Rhône-Alpes, Grenoble.

2004-2005   • Postdoc, Intelligent Autonomous Systems group, Informatics Institute, University of Amsterdam.

Professional Activities

Participation in Research Projects

2016-2018   • *Structured prediction for weakly supervised semantic segmentation*, funded by Facebook Artificial Intelligence Research (FAIR) Paris and French national research and technology agency (ANRT).

2015-2016   • *Incremental learning for object category localization*, funded by MBDA Systems.

2013-2016   • *Physionomie: Physiognomic Recognition for Forensic Investigation* , funded by French national research agency (ANR).

2011-2015   • *AXES: Access to Audiovisual Archives*, European integrated project, 7th Framework Programme.

2010-2013   • *Quaero Consortium for Multimodal Person Recognition*, funded by French national research agency (ANR).

2009-2012   • *Modeling multi-media documents for cross-media access*, funded by Xerox Research Centre Europe (XRCE) and French national research and technology agency (ANRT).

2008-2010   • *Interactive Image Search*, funded by French national research agency (ANR).

2006-2009   • *Cognitive-Level Annotation using Latent Statistical Structure (CLASS)*, funded by European Union Sixth Framework Programme.

2000-2005   • *Tools for Non-linear Data Analysis*, funded by Dutch Technology Foundation (STW).

Teaching

2015   • Lecturer in MSc course *Kernel Methods for Statistical Learning*, École Nationale Supérieure d'Informatique et de Mathématiques Appliquées (ENSIMAG), Grenoble, France.

2008-2015   • Lecturer in MSc course *Machine Learning and Category Representation*, École Nationale Supérieure d'Informatique et de Mathématiques Appliquées (ENSIMAG), Grenoble, France.

2003-2005   • Lecturer in MSc course *Machine learning: pattern recognition*, University of Amsterdam, The Netherlands.

Professional Activities (continued)

2003-2005	• Lecturer in graduate course *Advanced issues in neurocomputing*, Advanced School for Imaging and Computing, The Netherlands.
1997-2000	• Teaching assistant in courses MSc Artificial Intelligence, University of Amsterdam, The Netherlands.

Supervision of MSc and PhD Students

since 2016	• Pauline Luc, PhD, *Weakly supervised structured prediction for semantic segmentation.*
2016	• Thomas Lucas, MSc, *Recurrent neural network approaches for image captioning.*
2015	• Jerome Lesaint, MSc, *Image and video captioning.*
since 2013	• Shreyas Saxena, PhD, *Recognizing people in the wild.*
2013	• Shreyas Saxena, MSc, *Metric learning for face verification.*
2011-2015	• Dan Oneaţă, PhD, *Large-scale machine learning for video analysis.*
2010-2014	• Gokberk Cinbis, PhD, *Fisher kernel based models for image classification and object localization*, awarded AFRIF best thesis award 2014.
2009-2012	• Thomas Mensink, PhD, *Modeling multi-media documents for cross-media access*, awarded AFRIF best thesis award 2012.
2008-2011	• Josip Krapac, PhD, *Image search using combined text and image content.*
2006-2010	• Matthieu Guillaumin, PhD, *Learning models for visual recognition from weak supervision.*
2009	• Gaspard Jankowiak, intern, *Decision tree quantization of image patches for image categorization.*
2007-2008	• Thomas Mensink, intern, *Finding people in captioned news images.*
2005	• Markus Heukelom, MSc, *Face detection and pose estimation using part-based models.*
2003	• Jan Nunnink, MSc, *Large scale mixture modelling using a greedy expectation-maximisation algorithm.*
2003	• Noah Laith, MSc, *A fast greedy k-means algorithm.*

Associate Editor

since 2014	• International Journal of Computer Vision.
since 2011	• Image and Vision Computing Journal.

Chairs for International Conferences

• Tutorial Chair European Conference on Computer Vision: 2016.
• Area Chair IEEE Conference on Computer Vision and Pattern Recognition: 2015.
• Area Chair European Conference on Computer Vision: 2012, 2014.
• Area Chair British Machine Vision Conference: 2012, 2013, 2014.

Programme Committee Member for Conferences, including

• IEEE International Conference on Computer Vision: 2009, 2011, 2013, 2015.
• European Conference on Computer Vision: 2008, 2010, 2016.
• IEEE Conference on Computer Vision and Pattern Recognition: 2006–2014, 2016.
• Neural Information Processing Systems: 2006–2010, 2012–2013.
• Reconnaissance des Formes et l'Intelligence Artificielle: 2016.

Reviewer for International Journals, including

since 2008	• International Journal of Computer Vision.
since 2005	• IEEE Transactions on Neural Networks.
since 2004	• IEEE Transactions on Pattern Analysis and Machine Intelligence.

Reviewer of research grant proposals, including

2015	• Postdoctoral fellowship grant, Research Foundation Flanders (FWO)
2014	• Collaborative Research grant, Indo-French Centre for the Promotion of Advance Research (IFCPAR)
2010	• VENI grant, Netherlands Organisation for Scientific Research (NWO)

Miscellaneous

Research Visits

2011	• Visiting researcher Statistical Machine Learning group, NICTA Canberra, Autralia, May 2011.
2003	• Machine Learning group University of Toronto, Prof. Sam Roweis, Canada, May–September 2003.

Summer Schools & Workshops

2015   • DGA workshop on Big Data in Multimedia Information Processing, invited speaker, Paris, France, October 22.
- Physionomie workshop at European Academy of Forensic Science conference, co-organizer and speaker, Prague, Czech Republic, September 9.
- StatLearn workshop, invited speaker, April 13, 2015, Grenoble, France.

2014   • 3rd Croatian Computer Vision Workshop, Center of Excellence for Computer Vision, invited speaker, September 16, 2014, Zagreb, Croatia.

2011   • 2nd IST Workshop on Computer Vision and Machine Learning, Institute of Science and Technology, invited presentation, October 7, Vienna, Austria.
- Workshop on 3D and 2D Face Analysis and Recognition, Ecole Centrale de Lyon / Lyon University, invited presentation, January 28.

2010   • NIPS Workshop on Machine Learning for Next Generation Computer Vision Challenges, co-organizer, December 10, Whistler BC, Canada.
- ECCV Workshop on Face Detection: Where are we, and what next?, invited presentation, September 10, Hersonissos, Greece.
- INRIA Visual Recognition and Machine Learning Summer School, 1h lecture, July 26–30, Grenoble, France.

2009   • Workshop "Statistiques pour le traitement de l'image", Université Paris 1 Panthéon-Sorbonne, invited speaker, January 23.

2008   • International Workshop on Object Recognition, poster presentation, May 16–18 2008, Moltrasio, Italy.

Seminars

2015   • Société Francaise de Statistique, Institut Henri Poincaré, Paris, France, *Object detection with incomplete supervision*, October 23.
- Center for Machine Perception, Czech Technical University, Prague, Czech Republic, *Object detection with incomplete supervision*, September 8.
- Dept. of Information Engineering and Computer Science, University of Trento, Italy, *Object detection with incomplete supervision*, March 16.
- Computer Vision Center, Barcelona, Spain, *Object detection with incomplete supervision*, February 13.

2013   • Intelligent Systems Laboratory Amsterdam, University of Amsterdam, The Netherlands, *Segmentation Driven Object Detection with Fisher Vectors*, October 15.
- Media Integration and Communication Center at the University of Florence, Italy, *Segmentation Driven Object Detection with Fisher Vectors*, September 24.
- DGA workshop on Multimedia Information Processing (TIM 2013), Paris, France, *Face verification "in the wild"*, July 2.

2012   • Computer Vision and Machine Learning group, Institute of Science and Technology, Vienna, Austria, *Image categorization using Fisher kernels of non-iid image models*, June 11.
- Computer Vision Center, Barcelona, Spain, *Image categorization using Fisher kernels of non-iid image models*, June 4.
- TEXMEX Team, INRIA, Rennes, France, *Image categorization using Fisher kernels of non-iid image models*, April 20.

2011   • Statistical Machine Learning group, NICTA, Canberra, Australia, *Modelling spatial layout for image classification*, May 26.
- Canon Information Systems Research Australia, Sydney, Australia, *Learning structured prediction models for interactive image labeling*, May 20.

2010   • Laboratoire TIMC-IMAG, Learning: Models and Algorithms team, Grenoble, *Metric learning approaches for image annotation and face verification*, October 7.
- University of Oxford, Visual Geometry Group, Oxford, *TagProp: a discriminatively trained nearest neighbor model for image auto-annotation*, February 1.

2009   • Laboratoire Jean Kuntzmann, Grenoble, *Machine learning for semantic image interpretation*, June 11.
- University of Amsterdam, Intelligent Systems Laboratory, *Discriminative learning of nearest-neighbor models for image auto-annotation*, April 28.
- Université de Caen, Laboratoire GREYC, *Improving People Search Using Query Expansions*, February 5.

2008   • Computer Vision Center, Autonomous University of Barcelona, *Improving People Search Using Query Expansions*, September 26.
- Computer Vision Lab, Max Planck institute for Biological Cybernetics, *Scene Segmentation with CRFs Learned from Partially Labeled Images*, July 31.
- Textual and Visual Pattern Analysis team, Xerox Research Centre Europe, *Scene Segmentation with CRFs Learned from Partially Labeled Images*, April 24.

2006	• Parole group, LORIA Nancy, *Unsupervised learning of low-dimensional structure in high-dimensional data.*
	• Content Analysis group, Xerox Research Centre Europe, *Manifold learning: unsupervised, correspondences, and semi-supervised.*
2005	• Learning and Recognition in Vision group, INRIA Rhône-Alpes, *Manifold learning & image segmentation.*
	• Computer Engineering Group, Bielefeld University, *Manifold learning with local linear models and Gaussian fields.*
2004	• Algorithms and Complexity group, Dutch Center for Mathematics and Computer Science, *Semi-supervised dimension reduction through smoothing on graphs.*
2003	• Machine Learning team, Radboud University Nijmegen, *Spectral methods for dimension reduction and non-linear CCA.*
2002	• Information and Language Processing Systems group, University of Amsterdam, *A generative model for the Self-Organizing Map.*

Publications

In peer reviewed international journals

2016	• G. Cinbis, J. Verbeek, C. Schmid. *Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning.* IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear, 2016.
2015	• G. Cinbis, J. Verbeek, C. Schmid. *Approximate Fisher kernels of non-iid image models for image categorization.* IEEE Transactions on Pattern Analysis and Machine Intelligence, to appear, 2015.
	• H. Wang, D. Oneaţă, J. Verbeek, C. Schmid. *A robust and efficient video representation for action recognition.* International Journal of Computer Vision, to appear, 2015.
	• M. Douze, J. Revaud, J. Verbeek, H. Jégou, C. Schmid. *Circulant temporal encoding for video retrieval and temporal alignment.* International Journal of Computer Vision, to appear, 2015.
2013	• J. Sánchez, F. Perronnin, T. Mensink, J. Verbeek. *Image classification with the Fisher vector: theory and practice.* International Journal of Computer Vision 105 (3), pp. 222–245, 2013.
	• T. Mensink, J. Verbeek, F. Perronnin, G. Csurka. *Distance-based image classification: generalizing to new classes at near-zero cost.* IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (11), pp. 2624–2637, 2013.
	• T. Mensink, J. Verbeek, G. Csurka. *Tree-structured CRF models for interactive image labeling.* IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2), pp. 476–489, 2013.
2012	• M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid. *Face recognition from caption-based supervision.* International Journal of Computer Vision, 96(1), pp. 64–82, January 2012.
2010	• H. Jégou, C. Schmid, H. Harzallah, and J. Verbeek. *Accurate image search using the contextual dissimilarity measure.* IEEE Transactions on Pattern Analysis and Machine Intelligence 32(1), pp. 2–11, January 2010.
	• D. Larlus, J. Verbeek, F. Jurie. *Category level object segmentation by combining bag-of-words models with Dirichlet processes and random fields.* International Journal of Computer Vision 88(2), pp. 238–253, June 2010.
2009	• J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. *Learning color names for real-world applications.* IEEE Transactions on Image Processing 18(7), pp. 1512–1523, July 2009.
2006	• J. Verbeek, J. Nunnink, and N. Vlassis. *Accelerated EM-based clustering of large data sets.* Data Mining and Knowledge Discovery 13(3), pp. 291–307, November 2006.
	• J. Verbeek and N. Vlassis. *Gaussian fields for semi-supervised regression and correspondence learning.* Pattern Recognition 39(10), pp. 1864–1875, October 2006.
	• J. Verbeek. *Learning nonlinear image manifolds by global alignment of local linear models.* IEEE Transactions on Pattern Analysis and Machine Intelligence 28(8), pp. 1236–1250, August 2006.
2005	• J. Porta, J. Verbeek, B. Kröse. *Active appearance-based robot localization using stereo vision.* Autonomous Robots 18(1), pp. 59–80, January 2005.
	• J. Verbeek, N. Vlassis, and B. Kröse. *Self-organizing mixture models.* Neurocomputing 63, pp. 99–123, January, 2005.
2003	• J. Verbeek, N. Vlassis, and B. Kröse. *Efficient greedy learning of Gaussian mixture models.* Neural Computation 15(2), pp. 469–485, February 2003.
	• A. Likas, N. Vlassis, and J. Verbeek. *The global k-means clustering algorithm.* Pattern Recognition 36(2), pp. 451–461, February 2003.
2002	• J. Verbeek, N. Vlassis, and B. Kröse. *A k-segments algorithm for finding principal curves.* Pattern Recognition Letters 23(8), pp. 1009–1017, June 2002.

In peer reviewed international conferences

2014	• D. Oneaţă, J. Revaud, J. Verbeek, C. Schmid. *Spatio-Temporal Object Detection Proposals.* Proceedings European Conference on Computer Vision, September 2014.

Publications (continued)

- G. Cinbis, J. Verbeek, C. Schmid. *Multi-fold MIL Training for Weakly Supervised Object Localization.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2014.
- D. Oneață, J. Verbeek, C. Schmid. *Efficient Action Localization with Approximately Normalized Fisher Vectors.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2014.

2013
- G. Cinbis, J. Verbeek, C. Schmid. *Segmentation Driven Object Detection with Fisher Vectors.* Proceedings IEEE International Conference on Computer Vision, December 2013.
- D. Oneață, J. Verbeek, C. Schmid. *Action and Event Recognition with Fisher Vectors on a Compact Feature Set.* Proceedings IEEE International Conference on Computer Vision, December 2013.

2012
- T. Mensink, J. Verbeek, F. Perronnin, G. Csurka. *Metric learning for large scale image classification: generalizing to new classes at near-zero cost.* Proceedings European Conference on Computer Vision, October 2012. (oral)
- G. Cinbis, J. Verbeek, C. Schmid. *Image categorization using Fisher kernels of non-iid image models.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2012.

2011
- J. Krapac, J. Verbeek, F. Jurie. *Modeling spatial layout with Fisher vectors for image categorization.* Proceedings IEEE International Conference on Computer Vision, November 2011.
- G. Cinbis, J. Verbeek, C. Schmid. *Unsupervised metric learning for face identification in TV video.* Proceedings IEEE International Conference on Computer Vision, November 2011.
- J. Krapac, J. Verbeek, F. Jurie. *Learning tree-structured descriptor quantizers for image categorization.* Proceedings British Machine Vision Conference, September 2011.
- T. Mensink, J. Verbeek, G. Csurka. *Learning structured prediction models for interactive image labeling.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2011.

2010
- M. Guillaumin, J. Verbeek, C. Schmid. *Multiple instance metric learning from automatically labeled bags of faces.* Proceedings European Conference on Computer Vision, September 2010.
- M. Guillaumin, J. Verbeek, C. Schmid. *Multimodal semi-supervised learning for image classication.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2010. (oral)
- J. Krapac, M. Allan, J. Verbeek, F. Jurie. *Improving web image search results using query-relative classifiers.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, June 2010.
- T. Mensink, J. Verbeek, G. Csurka. *Trans Media Relevance Feedback for Image Autoannotation.* Proceedings British Machine Vision Conference, September 2010.
- T. Mensink, J. Verbeek, H. Kappen. *EP for efficient stochastic control with obstacles.* Proceedings European Conference on Artificial Intelligence, August 2010. (oral)
- J. Verbeek, M. Guillaumin, T. Mensink, C. Schmid. *Image Annotation with TagProp on the MIRFLICKR set.* Proceedings ACM International Conference on Multimedia Information Retrieval, March 2010. (invited paper)

2009
- M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid. *TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation.* Proceedings IEEE International Conference on Computer Vision, September 2009. (oral)
- M. Guillaumin, J. Verbeek, C. Schmid. *Is that you? Metric learning approaches for face identification.* Proceedings IEEE International Conference on Computer Vision, September 2009.
- M. Allan, J. Verbeek *Ranking user-annotated images for multiple query terms.* Proceedings British Machine Vision Conference, September 2009.

2008
- M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid. *Automatic face naming with caption-based supervision.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, June 2008.
- T. Mensink, and J. Verbeek. *Improving people search using query expansions: How friends help to find people.* Proceedings European Conference on Computer Vision, pp. 86–99, October 2008. (oral)
- J. Verbeek and B. Triggs. *Scene segmentation with CRFs learned from partially labeled images.* Advances in Neural Information Processing Systems 20, pp. 1553–1560, January 2008. (oral)
- H. Cevikalp, J. Verbeek, F. Jurie, and A. Kläser. *Semi-supervised dimensionality reduction using pairwise equivalence constraints.* Proceedings International Conference on Computer Vision Theory and Applications, pp. 489–496, January 2008.

2007
- J. van de Weijer, C. Schmid, and J. Verbeek. *Learning color names from real-world images.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, June 2007.
- J. Verbeek and B. Triggs. *Region classification with Markov field aspect models.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, June 2007.
- J. van de Weijer, C. Schmid, and J. Verbeek. *Using high-level visual information for color constancy.* Proceedings IEEE International Conference on Computer Vision, pp. 1–8, October 2007.

2006
- Z. Zivkovic and J. Verbeek. *Transformation invariant component analysis for binary images.* Proceedings IEEE Conference on Computer Vision and Pattern Recognition, pp. 254–259, June 2006.

2004
- J. Verbeek, S. Roweis, and N. Vlassis. *Non-linear CCA and PCA by alignment of local models.* Advances in Neural Information Processing Systems 16, pp. 297–304, January 2004. (oral)

2003
- J. Porta, J. Verbeek, and B. Kröse. *Enhancing appearance-based robot localization using non-dense disparity maps.* Proceedings International Conference on Intelligent Robots and Systems, pp. 980–985, October 2003.

- J. Verbeek, N. Vlassis, and B. Kröse. *Self-organization by optimizing free-energy.* Proceedings 11th European Symposium on Artificial Neural Networks, pp. 125–130, April 2003.

2002
- J. Verbeek, N. Vlassis, and B. Kröse. *Coordinating principal component analyzers.* Proceedings International Conference on Artificial Neural Networks, pp. 914–919, August 2002. (oral)
- J. Verbeek, N. Vlassis, and B. Kröse. *Fast nonlinear dimensionality reduction with topology preserving networks.* Proceedings 10th European Symposium on Artificial Neural Networks, pp. 193–198, April 2002. (oral)

2001
- J. Verbeek, N. Vlassis, and B. Kröse. *A soft k-segments algorithm for principal curves.* Proceedings International Conference on Artificial Neural Networks, pp. 450–456, August 2001.

Book chapters

2013
- T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. *Large scale metric learning for distance-based image classification on open ended data sets.* In: G. Farinella, S. Battiato, and R. Cipolla. *Advances in Computer Vision and Pattern Recognition*, Springer, 2013.

2012
- R. Benavente, J. van de Weijer, M. Vanrell, C. Schmid, R. Baldrich, J. Verbeek, and D. Larlus. *Color Names.* In: T. Gevers, A. Gijsenij, J. van de Weijer, and J. Geusebroek. *Color in Computer Vision*, Wiley, 2012.

Workshops and regional conferences

2015
- S. Saxena, and J. Verbeek. *Coordinated Local Metric Learning.* ICCV ChaLearn Looking at People workshop, December 2015.
- V. Zadrija, J. Krapac, J. Verbeek, and S. Šegvić. *Patch-level Spatial Layout for Classification and Weakly Supervised Localization.* German Conference on Pattern Recognition, October 2015.

2014
- M. Douze, D. Oneata, M. Paulin, C. Leray, N. Chesneau, D. Potapov, J. Verbeek, K. Alahari, Z. Harchaoui, L. Lamel, J.-L. Gauvain, C. Schmidt, and C. Schmid. *The INRIA-LIM-VocR and AXES submissions to Trecvid 2014 Multimedia Event Detection.* TRECVID Workshop, November, 2014.

2013
- R. Aly, R. Arandjelovic, K. Chatfield, M. Douze, B. Fernando, Z. Harchaoui, K. Mcguiness, N. O'Connor, D. Oneaţă, O. Parkhi, D. Potapov, J. Revaud, C. Schmid, J.-L. Schwenninger, D. Scott, T. Tuytelaars, J. Verbeek, H. Wang, and A. Zisserman. *The AXES submissions at TrecVid 2013.* TRECVID Workshop, November, 2013.
- H. Bredin, J. Poignant, G. Fortier, M. Tapaswi, V.-B. Le, A. Roy, C. Barras, S. Rosset, A. Sarkar, Q. Yang, H. Gao, A. Mignon, J. Verbeek, L. Besacier, G. Quénot, H. Ekenel, and R. Stiefelhagen. *QCompere @ REPERE 2013.* Workshop on Speech, Language and Audio for Multimedia, August 2013.

2012
- D. Oneaţă, M. Douze, J. Revaud, J. Schwenninger, D. Potapov, H. Wang, Z. Harchaoui, J. Verbeek, C. Schmid, R. Aly, K. Mcguiness S. Chen, N. O'Connor, K. Chatfield, O. Parkhi, and R. Arandjelovic, A. Zisserman, F. Basura, and T. Tuytelaars. *AXES at TRECVid 2012: KIS, INS, and MED.* TRECVID Workshop, November, 2012.
- H. Bredin, J. Poignant, M. Tapaswi, G. Fortier, V. Bac Le, T. Napoleon, H. Gao, C. Barras, S. Rosset, L. Besacier, J. Verbeek, G. Quénot, F. Jurie, H. Kemal Ekenel. *Fusion of speech, faces and text for person identification in TV broadcast.* ECCV Workshop on Information fusion in Computer Vision for Concept Recognition, October, 2012.

2011
- T. Mensink, J. Verbeek, and T. Caetano. *Learning to Rank and Quadratic Assignment.* NIPS Workshop on Discrete Optimization in Machine Learning, December 2011.

2010
- T. Mensink, G. Csurka, F. Perronnin, J. Sánchez, and J. Verbeek. *LEAR and XRCEs participation to Visual Concept Detection Task - ImageCLEF 2010.* Working Notes for the CLEF 2010 Workshop, September 2010.
- M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. *Apprentissage de distance pour l'annotation d'images par plus proches voisins.* Reconnaissance des Formes et Intelligence Artificielle, January 2010.

2009
- M. Douze, M. Guillaumin, T. Mensink, C. Schmid, and J. Verbeek. *INRIA-LEARs participation to ImageCLEF 2009.* Working Notes for the CLEF 2009 Workshop, September 2009.

2004
- J. Nunnink, J. Verbeek, and N. Vlassis. *Accelerated greedy mixture learning.* Proceedings Annual Machine Learning Conference of Belgium and the Netherlands, pp. 80–86, January 2004.

2003
- J. Verbeek, N. Vlassis, and J. Nunnink. *A variational EM algorithm for large-scale mixture modeling.* Proceedings Conference of the Advanced School for Computing and Imaging, pp. 136–143, June 2003.
- J. Verbeek, N. Vlassis, and B. Kröse. *Non-linear feature extraction by the coordination of mixture models.* Proceedings Conference of the Advanced School for Computing and Imaging, pp. 287–293, June 2003.

2002
- J. Verbeek, N. Vlassis, and B. Kröse. *Locally linear generative topographic mapping.* Proceedings Annual Machine Learning Conference of Belgium and the Netherlands, pp. 79–86, December 2002.

2001
- J. Verbeek, N. Vlassis, and B. Kröse. *Efficient greedy learning of Gaussian mixtures.* Proceedings 13th Belgian-Dutch Conference on Artificial Intelligence, pp. 251–258, October 2001.
- J. Verbeek, N. Vlassis, and B. Kröse. *Greedy Gaussian mixture learning for texture segmentation.* (oral) ICANN Workshop on Kernel and Subspace Methods for Computer Vision, pp. 37–46, August 2001.

2000 • J. Verbeek. *Supervised feature extraction for text categorization.* Proceedings Annual Machine Learning Conference of Belgium and the Netherlands, December 2000.

1999 • J. Verbeek. *Using a sample-dependent coding scheme for two-part MDL.* Proceedings Machine Learning & Applications (ACAI '99), July 1999.

Patents

2012 • T. Mensink, J. Verbeek, G. Csurka, and F. Perronnin. *Metric Learning for Nearest Class Mean Classifiers.* United States Patent Application 20140029839, Publication date: 01/30/2014, filing date: 07/30/2012, XEROX Corporation.

2011 • T. Mensink, J. Verbeek, and G. Csurka. *Learning Structured prediction models for interactive image labeling.* United States Patent Application 20120269436, Publication date: 25/10/2012, filing date: 20/04/2011, XEROX Corporation.

2010 • T. Mensink, J. Verbeek, and G. Csurka. *Retrieval systems and methods employing probabilistic cross-media relevance feedback.* United States Patent Application 20120054130, Publication date: 01/03/2012, filing date: 31/08/2010, XEROX Corporation.

Technical Reports

2013 • J. Sanchez, F. Perronnin, T. Mensink, J. Verbeek. *Image classification with the Fisher vector: theory and practice.* Technical Report RR-8209, INRIA, 2011.

2012 • T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. *Large scale metric learning for distance-based image classification.* Technical Report RR-8077, INRIA, 2011.

2011 • O. Yakhnenko, J. Verbeek, and C. Schmid. *Region-based image classification with a latent SVM model.* Technical Report RR-7665, INRIA, 2011.

• J. Krapac, J. Verbeek, F. Jurie. *Spatial Fisher vectors for image categorization.* Technical Report RR-7680, INRIA, 2011.

• T. Mensink, J. Verbeek, and G. Csurka. *Weighted transmedia relevance feedback for image retrieval and auto-annotation.* Technical Report RT-0415, INRIA, 2011.

2010 • M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. *Face recognition from caption-based supervision.* Technical Report RT-392, INRIA, 2010.

2008 • D. Larlus, J. Verbeek, and F. Jurie. *Category level object segmentation by combining bag-of-words models and Markov random fields.* Technical Report RR-6668, INRIA, 2008.

2005 • J. Verbeek, and N. Vlassis. *Semi-supervised learning with Gaussian fields.* Technical Report IAS-UVA-05-01, University of Amsterdam, 2005.

• J. Verbeek. *Rodent behavior annotation from video.* Technical Report IAS-UVA-05-02, University of Amsterdam, 2005.

2004 • J. Verbeek, and N. Vlassis. *Gaussian mixture learning from noisy data.* Technical Report IAS-UVA-04-01, University of Amsterdam, 2004.

2002 • J. Verbeek, N. Vlassis, and B. Kröse. *The generative self-organizing map: a probabilistic generalization of Kohonen's SOM.* Technical Report IAS-UVA-02-03, University of Amsterdam, 2002.

• J. Verbeek, N. Vlassis, and B. Kröse. *Procrustes analysis to coordinate mixtures of probabilistic principal component analyzers.* Technical Report IAS-UVA-02-01, University of Amsterdam, 2002.

2001 • A. Likas, N. Vlassis, and J. Verbeek. *The global k-means clustering algorithm.* Technical Report IAS-UVA-01-02, University of Amsterdam, 2001.

• J. Verbeek, N. Vlassis, and B. Kröse. *Efficient greedy learning of Gaussian mixtures.* Technical Report IAS-UVA-01-10, University of Amsterdam, 2001.

2000 • J. Verbeek, N. Vlassis, and B. Kröse. *A k-segments algorithm for finding principal curves.* Technical Report IAS-UVA-00-11, University of Amsterdam, 2000.