Pattern Recognition Letters
journal homepage: www.elsevier.com

Generative Part-Based Gabor Object Detector

Ekaterina Riabchenko[a], Joni-Kristian Kämäräinen[b,**]

[a] *Department of Mathematics and Physics, Lappeenranta University of Technology, Finland*
[b] *Department of Signal Processing, Tampere University of Technology, Finland*

ABSTRACT

Discriminative part-based models have become the approach for visual object detection. The models learn from a large number of positive and negative examples with annotated class labels and location (bounding box). In contrast, we propose a part-based generative model that learns from a small number of positive examples. This is achieved by utilizing "privileged information", sparse class-specific landmarks with semantic meaning. Our method uses bio-inspired complex-valued Gabor features to describe local parts. Gabor features are transformed to part probabilities by unsupervised Gaussian Mixture Model (GMM). GMM estimation is robustified for a small amount of data by a randomization procedure inspired by random forests. The GMM framework is also used to construct a probabilistic spatial model of part configurations. Our detector is invariant to translation, rotation and scaling. On part level invariance is achieved by pose quantization which is more efficient than previously proposed feature transformations. In the spatial model, invariance is achieved by mapping parts to an "aligned object space". Using a small number of positive examples our generative method performs comparably to the state-of-the-art discriminative method.

1. Introduction

Discriminative part-based models have become the approach for visual object detection and achieve state-of-the-art for various datasets, e.g., Caltech-101 (Fei-Fei et al., 2006), Caltech-256 (Griffin et al., 2007) and Pascal VOC (Everingham et al., 2010). Part-based models have two detection stages: detection of object parts and verifying detected parts' spatial configuration (constellation). The first methods with explicit spatial models were generative (Weber et al., 2000; Fei-Fei et al., 2007), but recent methods are based on discriminative learning from a large number of positive and negative examples with manually annotated class labels and location, e.g., the deformable part-based model (DPM) by Felzenszwalb et al. (2008, 2010).

The recent "big visual data" datasets, such as the ImageNet ILSVRC (Deng et al., 2009; Russakovsky et al., 2013), provide sufficient number of data for training deep architectures with millions of parameters to be optimized (Krizhevsky et al., 2012; Simonyan et al., 2013) and which are superior to the previous part-based models. However, with limited data and in specific applications part-based models and hybrids of deep architectures and part-based models perform extremely well (Wang et al., 2013b,a; Li et al., 2014). Despite dominance of discriminative learning in visual classification, generative models have desirable properties such as prior probabilities, learning from unlabelled data and visual synthesis, and therefore provide an alternative approach to be investigated.

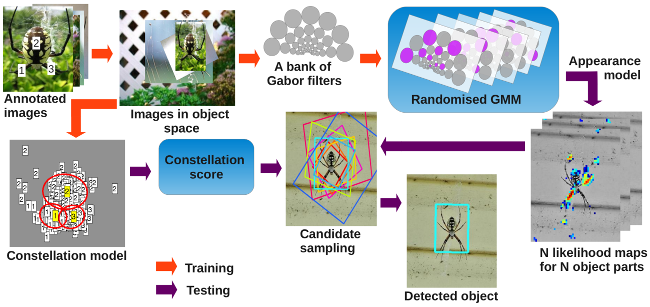

Fig. 1: Workflow of our generative learning and detection.

In this work, we propose a part-based generative model (Figure 1) that learns from a small number of positive examples. This is achieved by utilizing "privileged information", sparse

[**] Corresponding author. Tel. +358 50 3001851; `First.Second@tut.fi`

class-specific landmarks with semantic meaning. Our method uses bio-inspired complex-valued Gabor features to describe local parts. Gabor features are transformed to part probabilities by unsupervised Gaussian Mixture Model (GMM) probability densities. GMM estimation is robustified for a small amount of examples by novel randomized training inspired by random forests. GMMs are also used to represent spatial probabilities of the part configurations. Our detector is invariant to translation, rotation and scaling. On the part level, this is achieved by pose quantization which is more efficient than the previously proposed feature transformations (Kamarainen et al., 2006). In the spatial level, invariance is achieved by mapping parts to an "aligned object space". Using a small number of positive examples our generative method performs comparably to the state-of-the-art discriminative method.

2. Related work

Part-based models – The visual Bag-of-Words (BoW) (Sivic and Zisserman, 2003; Csurka et al., 2004) methods are omitted here since their spatial models are not explicit (see the recent survey by Huang et al. (2014)). The first part-based methods with constellation models were generative (Weber et al., 2000; Fei-Fei et al., 2007; Felzenszwalb and Huttenlockher, 2005), but since then the field has been dominated by discriminative learning and, in particular, the deformable part-based model (DPM) by Felzenszwalb et al. (2008, 2010). Recently, Xie et al. (2014) introduced a generative FRAME model which also uses Gabor features, but the model is computationally intensive, window-based and embeds geometry variation to appearance. Bourdev and Malik (2009) proposed a generative poselet model for part-based human pose detection, but its generality to other classes is unclear. Our method differs rather strongly from the above by the facts that it is generative, is generic, and has explicit models for the parts and constellation.

In particular, we extend our previous works of Gabor feature extraction (Kyrki et al., 2004) and Gaussian mixture model probabilistic part descriptor (Paalanen et al., 2006). Our quantized object pose space (Section 5.1) avoids the computationally expensive matrix shifts in (Kyrki et al., 2004) and the proposed randomized Gaussian mixture model (Section 4.3) can exploit a large Gabor filter bank and still learn a model from a few examples instead of hundreds required in (Paalanen et al., 2006). Preliminary results have been published in two conference papers, part detector in (Riabchenko et al., 2014b) and spatial model in (Riabchenko et al., 2014a), while this work refines the theory to form a single probabilistic framework, simplifies computation and improves performance by the quantized pose space, reports results from extensive experiments along with the full source code available in a public repository[1].

Contributions – We make the following contributions:

- A probabilistic (generative) local part descriptor using complex-valued multi-resolution Gabor features
    - In contrast to a small size Gabor bank used in the literature, we use a large bank and propose a method to identify a part-specific subset of the filters.
    - We avoid using heuristic prior distributions to learn from a small number of training examples by a novel random forest inspired generative learning procedure: *randomized Gaussian mixture model.*
    - We propose a likelihood-driven part detection procedure with efficient non-maximum suppression.
- A probabilistic part spatial constellation model in "aligned object space"
    - The model combines the probability terms of parts and their constellation.
    - The aligned space is formed by quantizing object appearances over rotation and scales.
- In extensive experiments on Caltech and ImageNet images our method performs favorably to the popular DPM.

3. Local Gabor descriptor

Gabor features have been successful in many vision applications such as iris and face recognition (Daugman, 1993; Shen and Bai, 2006). They are considered as texture descriptors (Bovik et al., 1990; Manjunath and Ma, 1996; Han and Ma, 2007), but local part description was one of the first applications (Lades et al., 1993; Wiskott et al., 1997). We adopt the multi-resolution Gabor feature - "simple Gabor feature space" - by Kyrki et al. (2004); Kamarainen et al. (2006):

$$\begin{aligned} \psi(x,y) &= \frac{f^2}{\pi\gamma\eta} e^{-(\frac{f^2}{\gamma^2}x'^2 + \frac{f^2}{\eta^2}y'^2)} e^{j2\pi f x'} \\ x' &= x\cos\theta + y\sin\theta \\ y' &= -x\sin\theta + y\cos\theta \ . \end{aligned} \tag{1}$$

f is the discrete tuning frequency, θ the rotation angle, γ the sharpness (bandwidth) of the major axis, and η of the minor axis. The spatial domain filter in (1) is a complex plane wave (a 2D Fourier basis function) multiplied by a Gaussian, and in the frequency domain it is a single real-valued Gaussian centered at f. The multi-resolution form and parametrisation in (1) enforces self-similarity: filters are scaled and rotated versions of each other, "Gabor wavelets".

Multi-resolution Gabor features are constructed from responses of filters tuned to the multiple frequencies f_m and orientations θ_n. Scales (f) are drawn from the exponential scale

$$f_m = k^{-m} f_{max}, \quad m = \{0, \ldots, M-1\} \tag{2}$$

where f_m is the mth frequency, $f_0 = f_{max}$ is the highest frequency, and $k > 1$ is the frequency scaling factor. The filter orientations are uniformly sampled:

$$\theta_n = \frac{n2\pi}{N}, \quad n = \{0, \ldots, N-1\} \tag{3}$$

where θ_n is the nth orientation and N is their total number.

The multi-resolution Gabor parameters f_{max}, k, M, N, γ and η are redundant and an intuitive parametrisation is to set the filter cross points to $p = 0.5$ when the filter envelopes cross at the

[1] https://bitbucket.org/EkaterinaRiabchenko/gabor_object_detector_code/

half magnitude providing sufficient “shiftability” (Sampo et al., 2006). In that case, the adjustable parameters are the highest frequency f_{max}, the number of frequencies m and the number of orientations n. The bandwidths γ and η are automatically set.

Descriptor invariance – The simple Gabor feature space part descriptor at the location (x_0, y_0) forms a Gabor response matrix:

$$\mathbf{G} = \begin{pmatrix} r(x_0,y_0;f_0,\theta_0) & r(x_0,y_0;f_0,\theta_1) & \cdots & r(x_0,y_0;f_0,\theta_{n-1}) \\ r(x_0,y_0;f_1,\theta_0) & r(x_0,y_0;f_1,\theta_1) & \cdots & r(x_0,y_0;f_1,\theta_{n-1}) \\ \vdots & \vdots & \ddots & \vdots \\ r(x_0,y_0;f_{m-1},\theta_0) & r(x_0,y_0;f_{m-1},\theta_1) & \cdots & r(x_0,y_0;f_{m-1},\theta_{n-1}) \end{pmatrix} \quad (4)$$

where rows denote different frequencies and columns orientations. The first row is the highest frequency $f_0 = f_{max}$ and the first column $\theta_0 = 0°$.

Column and row shifts of the response matrix provide invariance to geometric transformations, scaling and rotation (Kamarainen et al., 2006). For example, anti-clockwise rotation of an image by $\frac{\pi}{N}$ corresponds to a single shift operation:

$$\begin{pmatrix} r(x_0,y_0;f_0,\theta_{n-1})^* & r(x_0,y_0;f_0,\theta_0) & \Rightarrow & r(x_0,y_0;f_0,\theta_{n-2}) \\ r(x_0,y_0;f_1,\theta_{n-1})^* & r(x_0,y_0;f_1,\theta_0) & \Rightarrow & r(x_0,y_0;f_1,\theta_{n-2}) \\ \vdots & \vdots & \ddots & \vdots \\ r(x_0,y_0;f_{m-1},\theta_{n-1})^* & r(x_0,y_0;f_{m-1},\theta_0) & \Rightarrow & r(x_0,y_0;f_{m-1},\theta_{n-2}) \end{pmatrix} . \quad (5)$$

A similar shift operation exists for scaling, but in Section 5.1 we show that object poses are heavily “quantized” in the datasets and we propose invariant matching without the shift operations.

Importance of complex phase – Unlike the most other works which use only the magnitude information, our Gabor feature descriptor is complex-valued which is justified by the three important findings: 1) the phase information plays a dominant role for visual representation (Figure 2) (Oppenheim and Lim, 1981); 2) complex representation provides superior performance (see the experiments section); and 3) complex covariance matrix is more compact in our Gaussian mixture model probability density (Section 4.2).

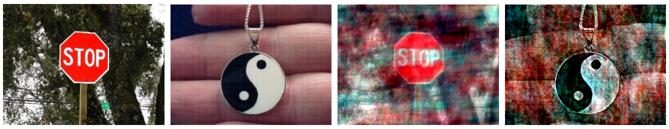

Fig. 2: Mixing the phase and magnitude information of two images. From left to right: two source images, reconstruction using yin yang magnitude and stop sign phase and using yin yang phase and stop sign magnitude.

4. Learning and detecting object parts

Our generative model builds upon the probabilistic models of object parts F_i, $p(\boldsymbol{G}|F_i)$, where $\boldsymbol{G}$ is the local Gabor descriptor computed at location (x_0, y_0). Our workflow has three processing stages (Fig. 1). Firstly, training images are transformed to an aligned object space where pose variation is removed (Sec. 4.1). Secondly, local Gabor part descriptors are computed at the part locations in the aligned space. Thirdly, generative probabilistic models of the parts are learned by unsupervised Gaussian mixture models (GMMs) (Sec. 4.2). We propose a randomized version of GMM, *Randomized GMM* (r-GMM), to learn part probability densities from a small number (tens) of examples (Sec. 4.3). The detection uses the same stages, but with estimated GMM pdfs and applies non-maximum suppression (Sec. 4.3.1). A preliminary version of the part detection was presented in (Riabchenko et al., 2014b), but here we show the complete class detection framework using the aligned object space and quantized poses (Sec. 5.1).

4.1. Aligned object space

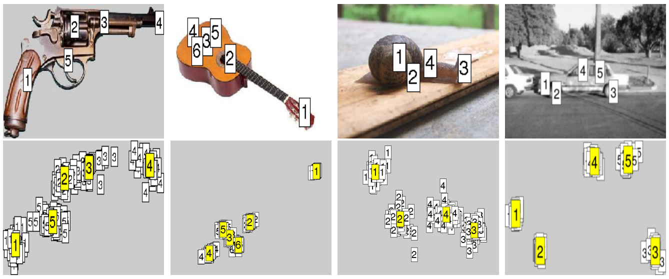

Fig. 3: Top: examples with annotated landmarks; bottom: all training images mapped to the “aligned object space” (see the text below).

Varying view point creates pose variance and if it is not taken into consideration, the local part descriptors will capture both appearance and pose changes which is undesirable. We wish to capture only appearance and therefore train the part detectors in an “aligned object space”. This can be achieved by fixing locations of some parts. M.C.Burl et al. (1998) and Hamouz et al. (2005) fixed two “anchor” parts, but that leads to two undesirable properties: variation of the anchors is transformed to other parts and the anchor parts become “golden” and must be always detected. Instead, we adopt the mean shape model by Cootes et al. (1995) and replace their approximate similarity (Procrustes algorithm) with the fast linear method of Umeyama (1991) (see the pseudo algorithm in Algorithm 1 and Figure 3). To remove outliers the method also stores a set of allowable transformations $\{H_{\text{prior}}\}$ to be used in Alg. 4. $\{H_{\text{prior}}\}$ is also used to calculate scale and orientation distributions of objects in a training set, which provide pose quantization (Sec. 5.1).

Algorithm 1 Aligned mean object space.

```
Select a seed example and use its parts' as the initial object space.
for all images do
    Estimate the mapping $H$ to the object space using the image parts (Umeyama, 1991)). //Store as $H_{prior}$ for Alg. 4.
    Transform object's parts to the aligned space.
    Refine by computing the average of transformed parts.
end for
Return the final aligned space and transformed images.
```

4.2. Part pdf using unsupervised Gaussian mixture models

The part model produces the part likelihood for a Gabor descriptor vector $\boldsymbol{g}_{D\times 1}$ which is constructed from (4) by concatenating the rows. The dimension D is the number of frequencies times the number of orientations: $D = M \times N$. We do not use only magnitudes as in the most Gabor works, but the full complex valued feature vector $\boldsymbol{g} \in \mathbb{C}^D$. That is justified by the fact that removing the phase destroys important information for detection. The complex multivariate Gaussian distribution is

$$\mathcal{N}(x;\mu,\Sigma) = \mathcal{N}^{\mathbb{C}}(x;\mu,\Sigma) = \frac{1}{\pi^D|\Sigma|}\exp\left[-(x-\mu)^*\Sigma^{-1}(x-\mu)\right] \tag{6}$$

where * denotes the adjoint matrix. The GMM extension is

$$p(x;\theta) = \sum_{c=1}^{C} \alpha_c\,\mathcal{N}(x;\mu_c,\Sigma_c) \tag{7}$$

where α_c is the weight of the cth component. The GMM pdf is defined by the parameters

$$p(\boldsymbol{G};\theta),\quad \theta = \{\alpha_1,\mu_1,\Sigma_1,\ldots,\alpha_C,\mu_C,\Sigma_C\}\ . \tag{8}$$

The complex representation also provides more robust estimation as the number of model parameters is reduced to

$$C(D^2+2D)+C-1 \tag{9}$$

from $C(2D^2+3D)+C-1$ (separate real and imaginary parts).

Since the number of components, C, is unknown we prefer unsupervised GMM methods. From the two popular methods, FJ by Figueiredo and Jain (2002) and greedy EM (GEM) by Verbeek et al. (2003), we found the FJ algorithm more accurate and stable. Additional robustness was implemented by enforcing estimated covariances to Hermitians on each iteration. More details about the robust GMM estimation can be found from (Paalanen et al., 2006).

4.3. *Optimal filter selection and randomized GMM*

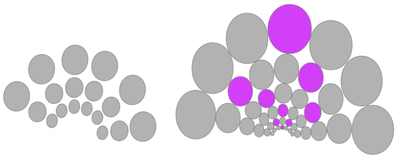

Fig. 4: The mainstream approach is to use all responses of a small filter bank (left) while in this work we use a large bank and a filter selection procedure.

Most works of Gabor features propose to use a small bank of Gabor filters, e.g., 4-6 orientations and 3-5 frequencies. This is not optimal and yields to two problems: redundant features due to correlation of the overlapping filters and over-fitting in the case of a limited number of training examples. In this work, we avoid these problems by using a large filter bank and selecting a subset of the filters for each part. The optimization procedure is based on feature randomization inspired by the highly successful random forest meta learning algorithm (Breiman, 2001). Using the bagging and randomization principles we extend the original unsupervised FJ GMM to *randomized FJ-GMM* (pseudo code in Alg. 2). All parameters of the randomized FJ GMM depend on the available computing time. To perform the full optimization within one hour we set the parameters to $N = 4$, $K = 9$, $B = 5$ and $T = 50$. However, the only parameter which needs to be set is N while all others can be optimized in the N-fold cross-validation loop.

Algorithm 2 Randomized FJ GMM (r-FJ-GMM).

```
N-fold randomization of the training data.
Compute full bank features for all parts.
for $T$ iterations do
    Randomly select $K$ filter responses from all training set images
    Run F-J GMM estimator using the selected features
    Evaluate landmark detection accuracy for the validation set
    Store the filter set and its performance
end for
Choose the $B$ best sets and use their combination as the GMM pdf
```

4.3.1. *Gabor part detector*

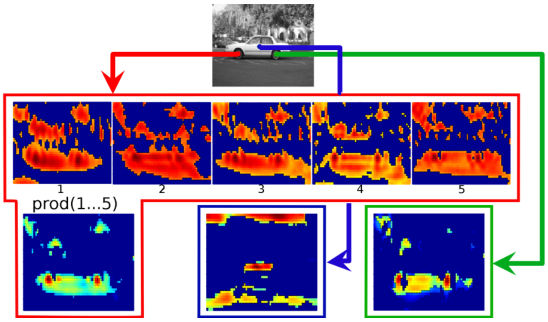

Fig. 5: Three car parts (denoted by red, green and blue) are detected from the input image (top). Middle row shows the random GMM probability maps of the front tyre. At the bottom are the fused final probability maps of each part.

The landmark detection becomes tricky with multiple (B) GMMs and we need a procedure to fuse the B likelihoods. The best result is achieved by a combination of thresholding, multiplication and non-maximum suppression (Alg. 3).

Algorithm 3 Single part detection using Gabor & r-FJ-GMM.

```
Require: $B$ GMMs of $K$ Gabor features.
Compute $B$ likelihood maps using the estimated GMMs.
Threshold likelihood maps to retain $P_1$% highest vals.
Multiply the $B$ thresholded maps.
Apply recursive global maximum search with suppression
```

A single GMM using a subset of Gabor features produces many false positives, but these are different (random) for B GMMs while the true positives are shared. The false positives can be effectively removed by the multiplication - the product rule in the combination theory (Kittler et al., 1998). To further sparsify the final likelihood map, we threshold each map to retain only $P_1 = 40\%$ highest likelihoods. This approach automatically adapts to “difficulty” of each part and provides high likelihoods for the correct locations (Fig. 5).

5. Spatial part constellation model

In Section 4 we introduced the part probabilities $p(\boldsymbol{G}, F_i) = p(\boldsymbol{g};\theta_i)$ and here we define the constellation probability over spatial locations of the parts $p(\boldsymbol{x};\theta_c)$ (c: class id). We define the constellation model also in the aligned space to make it invariant to pose changes. The part configuration can be represented

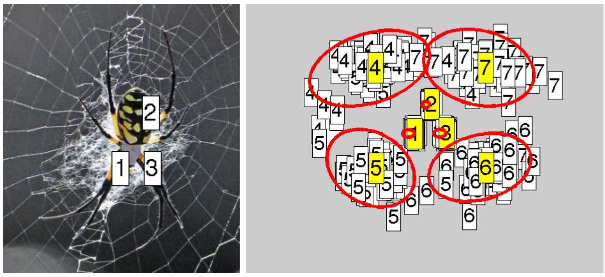

Fig. 6: Part distributions in the aligned object space. Note that the part probabilities (1-3) are by order of magnitude denser than the bounding box corners (4-7) which indicates inconsistent annotation.

by spatial Gaussians in the aligned space (Figure 6):

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2 \dots \boldsymbol{x}_N|\theta_c) = p(\boldsymbol{x}_1|\boldsymbol{\mu}_1, \Sigma_1) \cdot p(\boldsymbol{x}_2|\boldsymbol{\mu}_2, \Sigma_2) \cdot \dots p(\boldsymbol{x}_N|\boldsymbol{\mu}_N, \Sigma_N) \tag{10}$$

where N is the number of parts (3-6 depending on the class).

5.1. *Quantized pose space*

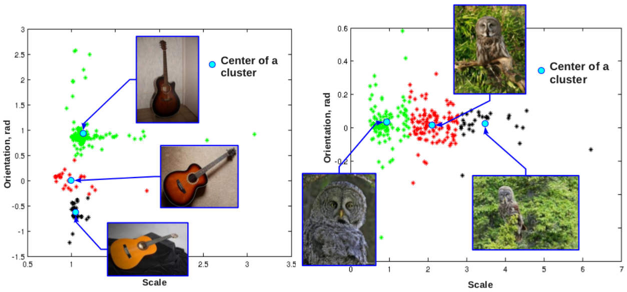

Fig. 7: Examples of "quantized poses" of ImageNet classes. For example, guitars are mainly quantized in orientation while owls in scale.

One reason why methods that do not take special care of pose changes still perform well with the existing benchmarks (Caltech, Pascal VOC, ImageNet) is that many objects are captured from a "quantized" set of viewpoints (Figure 7). This fact yields from the laws of physics, scene structures, and the way people prefer to capture images, e.g., pictures of sofas are usually frontal as their backs are turned towards walls. On the part level, invariance can be achieved by the matrix shifts exemplified in Section 3, but this is unnecessary if the pose "quantums" are known. The state-of-the-art discriminative method (DPM) by Felzenszwalb et al. (2010) exploits the quantized poses by clustering image bounding box ratios. However, the bounding box annotations are inconsistent (Fig. 6) and better results are achieved by using the part annotations - the "privileged information". We run c-means ($c = 3$) clustering in the parts' scale-orientation space, calculated from $\{H_{\text{prior}}\}$ in Alg. 1, and select the mean images as the aligned object space seeds. All examples can still be used to train each "quantized model" (Fig. 7) by transforming images to a specific cluster center "aligned object space".

5.2. *Maximum-likelihood object detection*

The part and constellation probabilities in Section 4 and 5.1 can be used in Bayesian detection, e.g., by Markov chain Monte-Carlo (MCMC) sampling, but this is computationally slow. Instead, we run exhaustive search to find multiple high likelihood candidates: $h_{ML} = \{hyp_{BEST}\}$. The exhaustive search is doable by limiting the number of best part candidates to 5-10 (Algorithm 4). The only parameter is the omission probability ($P_{landmark}$) that is set to cover 95% quantile in the training set.

Algorithm 4 Gabor object detection

```
Initialize the set of best hypotheses $\{hyp_{BEST}\}$ to $\emptyset$
for each combination of 2 different parts do
    Estimate similarity $H$ from the input image to the object space
    if $H \notin \{H_{\text{prior}}\}$ then
        Skip this hypothesis. // Not allowed transformation
    end if
    Transform detected parts to the object space using $H$
    For each part compute the spatial likelihood (Eq. 10).
    Select the highest likelihood parts (omit if below $P_{landmark}$).
    Compute the detection probability $p$ (Section 5.3)
    if $p$ is better than for any in $\{H_{best}\}$ then
        Estimate $H^{-1}$ back to the image space
        Transform the parts (use mean part for the omitted).
        Add hypothesis to $\{hyp_{BEST}\}$.
    end if
end for
Return the best hypotheses in $\{hyp_{BEST}\}$
```

5.3. *Maximum-likelihood probability p in Algorithm 4*

The Bayesian rule would require a background model – *X_not_at_that_location* probability – similar to (Fei-Fei et al., 2006) with an ad hoc model and negative examples. Instead, we use again the maximum likelihood estimate.

The full likelihood consists of the two elements, appearance and pose, which , by assuming their independence, can be written as (C denotes class number and i its part number);

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{g}|\theta_{c,i}, \theta_c) &= p(\boldsymbol{x}|\theta_c) \times p(\boldsymbol{g}, \theta_{c,i}) = \\ &\underbrace{p(\boldsymbol{x}_1|\boldsymbol{\mu}_1, \Sigma_1) \cdot \ldots \cdot p(\boldsymbol{x}_N|\boldsymbol{\mu}_N, \Sigma_N)}_{\text{constellation}} \underbrace{p(\boldsymbol{g}|\theta_{c,i})}_{\text{appearance}} . \end{aligned} \tag{11}$$

To simplify computation, we assume conditional independence for appearance, and to tolerate detection failures (occlusion) we include all combinations of 1, 2, ..., N-1, N detected parts which leads to 1st, 2nd, ..., Nth order probability terms:

$$\begin{cases} p(\boldsymbol{g}_1|\theta_{c,1}) + p(\boldsymbol{g}_2|\theta_{c,2}) + \ldots + p(\boldsymbol{g}_N|\theta_{c,N}) + \ldots \\ p(\boldsymbol{g}_1|\theta_{c,1})p(\boldsymbol{g}_2|\theta_{c,2}) + p(\boldsymbol{g}_1|\theta_{c,1})p(\boldsymbol{g}_3|\theta_{c,3}) + \\ \ldots \end{cases} \tag{12}$$

which requires summation of $\sum_{k=1\ldots N}\binom{N}{k}$ product terms. Good iterative implementations perform well up to 10-15 parts after which approximations (e.g., Stirling's formula) must be utilized. We used the exact method.

6. Experiments and examples

Data – We randomly selected classes for which somehow intuitive parts can be annotated from the Caltech-101 and ImageNet benchmarks. Object parts were annotated on rich textural regions like eyes in faces or tires in car side images. Images with rarely appearing 3D view points were removed despite the fact that separate models could be trained for them. All image classes were randomly divided into approximately equal sized training and test sets having 28 training and 24 test images in the smallest Caltech-101 class, dollar bill, while the biggest tested class - BioID faces, was represented with 507 training images and 1024 test images.

Performance evaluation – In the part detection experiments with BioID face images, we utilized the standard intra-ocular distance, d_{eye}, that measures errors of landmarks divided by the ground truth eye distance. The normalized distance ≤ 0.05 is considered as excellent, ≤ 0.10 good and > 0.25 as a failure.

The object detection accuracy is calculated according to the ImageNet protocol: the detection result is correct if the detected bounding box overlaps with the ratio $A \geq 0.5$. The ratio is calculated as $A = (BB_{gt} \cap BB_p)/(BB_{gt} \cup BB_p)$, where BB_{gt} is the area of a ground truth bounding box and BB_p of a predicted bounding box. Precision-recall curves are characterised by the average precision (AP) computed at 11 fixed points of the recall curve.

6.1. *BioID facial landmark detection*

Our part detector in Section 4 is a general method, but to show its performance quantitatively we report results for the well-known problem of facial landmark detection. The results are reported for the BioID database according to the BioID evaluation protocol. We compare our method to state-of-the-art detectors: LEAR by (Martinez et al., 2013), MK-SVM by (Rapp et al., 2011) and CLM by (Cristinacce and Cootes, 2008). The results are shown in Figure 8. It is noteworthy, that our part-detector without any special processing for facial parts performs favourably as compared to very dedicated facial landmark detection methods from the recent literature. Our detector basically misses some 10% of the most difficult landmarks. On the other hand, more than half of the landmarks per image are correctly found in 73% of images for the most strict metric ($\leq$ 0.05). For the less strict metrics more than 10 landmarks per image are found practically in 90 – 98% of the images. It should be noted that all other methods are discriminative, include special processing and a full facial landmark model while our method just returns the one best candidate of each landmark with no spatial regulation.

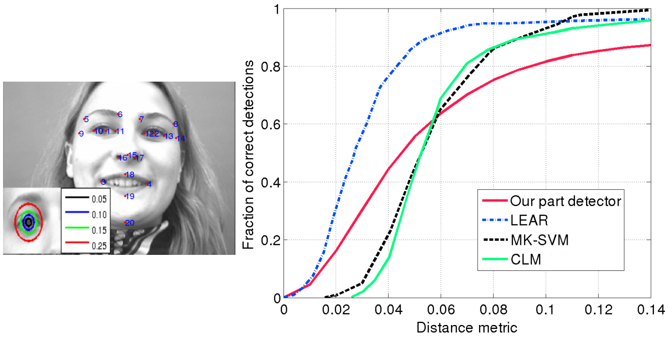

Fig. 8: BioID landmarks & the detection thresholds and the cumulative errors.

6.2. *Object part detection*

Similar to the facial landmarks we may annotate object parts and compute the detection error graphs (Figure 9). These results verify that detection of 5-10 highest likelihood part candidates provides good overall performance (at least 80% of the landmarks detected within the accuracy of 0.01). The detection likelihoods are illustrated in Figure 10.

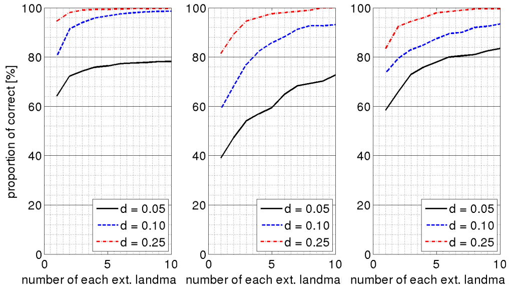

Fig. 9: Cumulative error graphs for the landmark detection of (from the left): motorbikes, revolvers and grand piano (Caltech).

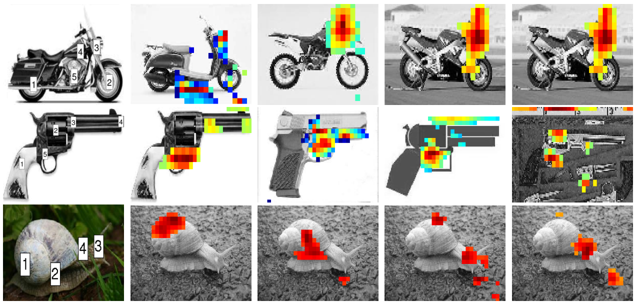

Fig. 10: Detection likelihoods. Left: examples with landmarks. The two top rows show likelihoods of a single landmark (#3 & #5) in multiple images and the bottom row likelihoods of multiple landmarks in the same image.

6.3. *Visual class detection*

Caltech – In the first experiment we used the Caltech101 classes for which 2D pose changes are minor and therefore represent "ideal case". We compare our results to the state-of-the-art discriminative method, the deformable part-based model (DPM) by Felzenszwalb et al. (2010). In addition, we construct "DPM-noneg" for which we use only one negative image to mimick positive only learning. The results are given in Table 1. Our method performs comparably to DPM-noneg being clearly superior for dragonfly and airplanes. For yin yang class generative method outperform DPM.

In Table 1 we provide results for using magnitude information of Gabor features (Our-nophase). The magnitude features perform well for many classes, but the results are always clearly inferior to the complex valued Gabor features used in our work.

To investigate our method's sensitivity to landmark selection, we tested its performance by replacing the manual landmarks with automatically selected landmarks. These landmarks are nine points set on a regular grid inside each bounding box. The results in Table 1 (Our-genLM) show that these landmarks perform surprisingly well and are significantly worse only for two classes, cars and airplanes. Unsupervised search of good object landmarks is an open problem for the future research.

Table 1: Caltech detection results (average precision).

	🚗	💵	STOP	🔫	🦋	🎹
DPM	100,0	100,0	100,0	99,8	100,0	100,0
DPM-noneg	99,6	89,1	99,7	97,4	60,2	87,4
Our	95,2	95,8	94,1	89,0	95,7	92,3
Our-nophase	69,5	92,9	66,5	74,4	94,1	67,6
Our-genLM	56,0	91,7	81,7	80,3	83,3	83,0
	🕎	☯	👩	⌚	✈	🏍
DPM	90,1	89,7	100,0	90,7	90,7	100,0
DPM-noneg	87,3	92,7	97,0	98,2	51,3	99,8
Our	81,7	93,1	99,6	97,2	86,7	96,9
Our-nophase	68,6	78,7	96,1	95,6	79,3	86,8
Our-genLM	78,4	100,0	98,4	82,2	67,7	87,4

Number of training examples – The main reason for the randomized GMM in Section 4.3 is to learn a detector from a few examples such as from only 28 positive examples of the dollar bill category in Table 1. In this experiment we investigated the generality of our method by keeping all parameters fixed, but using a various number of training examples. The average detection performances are shown in Figure 11 for four Caltech classes. These results verify that our generative detector succeeds to learn an object model already from only five positive examples and the performance steadily improves until approximately 20 examples after which the performance start to saturate for the selected set of landmarks.

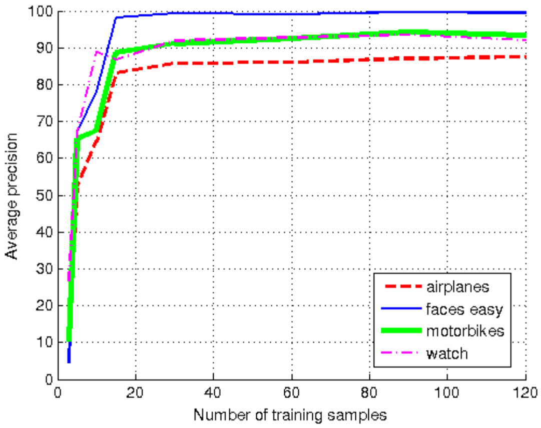

Fig. 11: Class detection performance as functions of the number of training images (average of ten random runs) for the Caltech airplanes, watch, faces and motorbikes categories.

ImageNet classes – ImageNet classes represent the most difficult dataset available and even for a few classes there is a clear drop in the DPM performance. DPM-noneg cannot learn any more part descriptors well and our method is clearly superior to the DPM-noneg (Table 2 and Figure 12). In addition to the plain bounding box detection our method also provides estimate of the object pose (Figure 13).

Table 2: ImageNet detection results (canonic: pose variance removed).

	🦉	🎸	🕷	🎹	🐌
DPM	90,9	90,7	88,0	90,5	86,8
DPM-noneg	76,4	86,2	24,6	39,0	20,8
Our	81,7	80,2	52,5	59,5	31,4
Our (canonic)	96,3	88,4	71,5	66,8	58,2

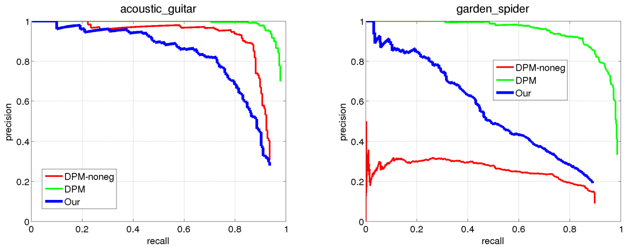

Fig. 12: ImageNet precision-recall curves.

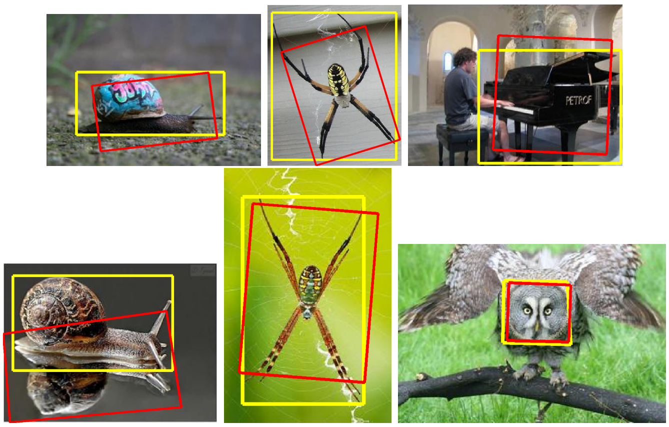

Fig. 13: Example ImageNet detections. Our method also provides poses. Yellow boxes - groundtruth, red boxes - our detections.

7. Conclusion

In contrast to the mainstream, we investigated the generative part-based approach for visual class detection and proposed a Gabor feature based framework for learning generative part-based models from a small number of positive examples. The framework for which source code is available performed favourably to the state-of-the-art discriminative method. We found that the part detection is the most essential stage for successful detection and in our future work we will replace the ad hoc manual part selection with part optimization process to push generative approaches even further.

References

Bourdev, L., Malik, J., 2009. Poselets: Body part detectors trained using 3D human pose annotations, in: Proc. of Int. Conf. on Computer Vision (ICCV), pp. 1365–1372.

Bovik, A.C., Clark, M., Geisler, W.S., 1990. Multichannel texture analysis using localized spatial filters. IEEE Trans. on Pattern Analysis and Machine Intelligence 12, 55–73.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Cootes, T., Taylor, C., Cooper, D., Graham, J., 1995. Active shape models – their training and application. Computer Vision and Image Understanding 61, 38–59.

Cristinacce, D., Cootes, T., 2008. Automatic feature localisation with constrained local models. Pattern Recognition 41, 3054–3067.

Csurka, G., Dance, C., Willamowski, J., Fan, L., Bray, C., 2004. Visual categorization with bags of keypoints, in: ECCV Workshop on Statistical Learning in Computer Vision, pp. 1–2.

Daugman, J., 1993. High confidence visual recognition of persons by a test of statistical independence. IEEE Trans. on Pattern Analysis and Machine Intelligence 15, 1148–1161.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database, in: Proc. of Computer Vision and Pattern Recognition (CVPR), pp. 248–255.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The Pascal visual object classes (VOC) challenge. Int. Journal of Computer Vision 88, 303–338.

Fei-Fei, L., Fergus, R., Perona, P., 2006. One-shot learning of object categories. IEEE Trans. on Pattern Analysis and Machine Intelligence 28, 594–611.

Fei-Fei, L., Fergus, R., Perona, P., 2007. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding 106, 59–70.

Felzenszwalb, P., Huttenlockher, D., 2005. Pictorial structures for object recognition. Int J of Comput Vis 61, 55–79.

Felzenszwalb, P., McAllester, D., Ramanan, D., 2008. A discriminatively trained, multiscale, deformable part model, in: Proc. of Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part based models. IEEE Trans. on Pattern Analysis and Machine Intelligence 32, 1627–1645.

Figueiredo, M., Jain, A., 2002. Unsupervised learning of finite mixture models. IEEE Trans. on Pattern Analysis and Machine Intelligence 24, 381–396.

Griffin, G., Holub, A., Perona, P., 2007. Caltech-256 Object Category Dataset. Technical Report 7694. California Institute of Technology.

Hamouz, M., Kittler, J., Kamarainen, J.K., Paalanen, P., Kalviainen, H., Matas, J., 2005. Feature-based affine-invariant localization of faces. IEEE Trans. on Pattern Analysis and Machine Intelligence 27, 1490–1495.

Han, J., Ma, K.K., 2007. Rotation-invariant and scale-invariant gabor features for texture image retrieval. Image and Vision Computing 25, 1474–1481.

Huang, Y., Wu, Z., Wang, L., Tan, T., 2014. Feature coding in image classification: A comprehensive study. IEEE Trans. on Pattern Analysis and Machine Intelligence 36, 493–506.

Kamarainen, J.K., Kyrki, V., Kälviäinen, H., 2006. Invariance properties of Gabor filter based features - overview and applications. IEEE Trans. on Image Processing 15, 1088–1099.

Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., 1998. On combining classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence 20, 226–239.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. ImageNet classification with deep convolutional neural networks, in: Proc. of Neural Information Processing Systems (NIPS), pp. 1106–1114.

Kyrki, V., Kamarainen, J.K., Kälviäinen, H., 2004. Simple Gabor feature space for invariant object recognition. Pattern Recognition Letters 25, 311–318.

Lades, M., Vorbrüggen, J.C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R.P., Konen, W., 1993. Distortion invariant object recognition in the dynamic link architecture. IEEE Trans. on Computers 42, 300–311.

Li, L., Su, H., Lim, Y., Fei-Fei, L., 2014. Object bank: An object-level image representation for high-level visual recognition. Int. Journal of Computer Vision 107, 20–39.

Manjunath, B., Ma, W., 1996. Texture features for browsing and retrieval of image data. IEEE Trans. on Pattern Analysis and Machine Intelligence 18, 837–842.

Martinez, B., Valstar, M., Binefa, X., Pantic, M., 2013. Local evidence aggregation for regression based facial point detection. IEEE Trans. on Pattern Analysis and Machine Intelligence 35, 1149–1163.

M.C.Burl, M.Weber, P.Perona, 1998. A probabilistic approach to object recognition using local photometry and global geometry, in: Proc. of European Conference on Computer Vision (ECCV), pp. 628–641.

Oppenheim, A., Lim, J., 1981. The importance of phase in signals. Proceedings of the IEEE 69, 529–541.

Paalanen, P., Kamarainen, J.K., Ilonen, J., Kälviäinen, H., 2006. Feature representation and discrimination based on Gaussian mixture model probability densities - practices and algorithms. Pattern Recognition 39, 1346–1358.

Rapp, V., Senechal, T., Bailly, K., Prevost, L., 2011. Multiple kernel learning SVM and statistical validation for facial landmark detection, in: Proc. of Int. Conf. on Automatic Face and Gesture Recognition (FG), pp. 265–271.

Riabchenko, E., Kämäräinen, J.K., Chen, K., 2014a. Density-aware part-based object detection with positive examples, in: Proc. of Int. Conf. on Pattern Recognition (ICPR), pp. 2814–2819.

Riabchenko, E., Kämäräinen, J.K., Chen, K., 2014b. Learning generative models of object parts from a few positive examples, in: Proc. of Int. Conf. on Pattern Recognition (ICPR), pp. 2287–2292.

Russakovsky, O., Deng, J., Krause, J., Berg, A., Fei-Fei, L., 2013. Imagenet Large Scale Visual Recognition Challenge 2013 (ILSVRC2013) results. http://www.image-net.org/challenges/LSVRC/2013/results.php.

Sampo, J., Kamarainen, J.K., Heiliö, M., Kälviäinen, H., 2006. Measuring translation shiftability of frames. Computers & Mathematics with Applications 52, 1089–1098.

Shen, L., Bai, L., 2006. A review on Gabor wavelets for face recognition. Pattern Analysis and Applications 9, 273–292.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep fisher networks for large-scale image classification, in: Proc. of Neural Information Processing Systems (NIPS), pp. 163–171.

Sivic, J., Zisserman, A., 2003. Video Google: A text retrieval approach to object matching in videos, in: Proc. of Int. Conf. on Computer Vision (ICCV), pp. 1470–1477.

Umeyama, S., 1991. Least-squares estimation of transformation parameters between two point patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence 13, 376–380.

Verbeek, J.J., Vlassis, N., Kröse, B., 2003. Efficient greedy learning of Gaussian mixture models. Neural Computation 5, 469–485.

Wang, X., Lin, L., Huang, L., Yan, S., 2013a. Incorporating structural alternatives and sharing into hierarchy for multiclass object recognition and detection, in: Proc. of Computer Vision and Pattern Recognition (CVPR), pp. 3334–3341.

Wang, X., Yang, M., Zhu, S., Lin, Y., 2013b. Regionlets for generic object detection, in: Proc. of Int. Conf. on Computer Vision (ICCV), pp. 17–24.

Weber, M., Welling, M., Perona, P., 2000. Unsupervised learning of models for recognition, in: Proc. of European Conference on Computer Vision (ECCV), pp. 18–32.

Wiskott, L., Fellous, J.M., Krüger, N., von der Malsburg, C., 1997. Face recognition by elastic bunch graph matching. IEEE Trans. on Pattern Analysis and Machine Intelligence 19, 775–779.

Xie, J., Hu, W., Zhu, S.C., Wu, Y., 2014. Learning inhomogeneous FRAME models for object patterns, in: Proc. of Computer Vision and Pattern Recognition (CVPR), pp. 1035–1042.