TENSOR-VARIATE GAUSSIAN PROCESSES REGRESSION AND ITS APPLICATION TO VIDEO SURVEILLANCE

Qibin Zhao[1]*, Guoxu Zhou*[1]*, Liqing Zhang*[2]*, Andrzej Cichocki*[1]

[1] RIKEN Brain Science Institute, Japan, [2] Shanghai Jiao Tong University, China

ABSTRACT

We present a novel framework for tensor valued Gaussian processes (GP) regression, which exploits a covariance function defined on tensor representation of data inputs. In this way, we bring together the powerful GP methods supported by Bayesian inference and higher-order tensor analysis techniques into one framework. This enables us to account for the underlying structure of data within the model, providing a powerful framework for structural data analysis, such as 3D video sequences. To this end, we propose a new kernel function with tensor arguments under the assumption of generative models, in the form of product kernels where a symmetrical Kullback-Leibler divergence measure is exploited to define the covariance function for tensorial data. A fully Bayesian treatment is employed to estimate the hyperparameters and infer the predictive distributions. Simulation results on both the synthetic data and a real world application of estimating the crowd size from 3D videos demonstrate the effectiveness of the proposed framework.

Index Terms— Tensor, tensor kernel, Gaussian processes

1. INTRODUCTION

Gaussian processes (GP) are a class of probabilistic models specifying a distribution over function spaces, whereby the inference is directly performed in the function space. Due to those desirable properties, GPs have gained much attention in recent years; in addition the prediction based on GP models often takes the form of a full predictive distribution [1]. This makes them powerful tools for Bayesian nonlinear and non-parametric regression, in which the prior distributions over latent function can be defined implicitly by the mean and covariance function. The hierarchical Bayesian modeling with GPs and inference conducted in the function space by evaluating the posterior process are presented in [2]. However, the most existing aspects of GPs can only be achieved in multivariate input data spaces.

Tensors (also called multiway arrays) are a generalization of vectors and matrices to higher dimensions, and are equipped with corresponding multilinear operators. The theory and algorithms of tensor decomposition (or factorization techniques), which can be regarded as the multilinear generalization of constrained matrix factorizations, have attracted much interest in the past decade, see e.g. [3, 4, 5], and have been successfully applied to problems in unsupervised learning or exploratory data analysis. Multiway structures typically allow us to capture the structure of the data, usually from *a priori* information about original data nature. This promises advantages over matrix factorizations, owing to a more effective use of the underlying properties of the structured data.

In order to combine the powerful GP model and Bayesian inference with tensor representation for structured data, in this study, we investigate the GP regression model based on tensor-variate inputs. In Sec. 2, we introduce a hierarchical Bayesian models for tensor-based GPs and the fundamental inference of predictive distribution. The crucial issue is the covariance function for tensorial inputs, governing a prior distribution over the latent function, which is defined based on multi-mode product kernels and probabilistic generative models in Sec. 3. Subsequently, the hyperparameter learning procedure is described in Sec. 4. The effectiveness of the proposed model and the corresponding inference is demonstrated by simulations on both synthetic data and a real-world application of counting pedestrians from a crowd based on 3D video sequences in Sec. 5. Sec. 6 concludes the study.

2. MODEL AND INFERENCE

Given a paired dataset of N observations $\mathcal{D} = \{(\mathcal{X}_n, y_n)|n = 1, \ldots, N\}$, the tensor inputs for all N cases are aggregated in an $M+1$th-order *design tensor* $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N \times I_1 \times \cdots \times I_M}$, and the targets are collected in the vector $\mathbf{y} = [y_1, \ldots, y_N]^T$. After observing the training data $\mathcal{D} = \{\boldsymbol{\mathcal{X}}, \mathbf{y}\}$, we are interested in making inferences about the relationship between inputs and targets, i.e., the conditional distribution of the targets given the inputs, and in making prediction for a new input $\mathcal{X}_*$ that we have not seen in the training set. The distribution of observations can be factored over cases in the training set by $\mathbf{y} \sim \prod_{n=1}^{N} \mathcal{N}(y_n|f_n, \sigma^2)$, where f_n denotes latent function $f(\mathcal{X}_n)$. A Gaussian process prior can be placed over the latent function, which implies that any finite subset of latent variables has a multivariate Gaussian distribution, denoted by

$$f(\mathcal{X}) \sim \mathcal{GP}(m(\mathcal{X}), k(\mathcal{X}, \mathcal{X}') \mid \boldsymbol{\theta}) \tag{1}$$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

where $m(\mathcal{X})$ is mean function and usually is set to zero for notational simplicity, and $k(\mathcal{X}, \mathcal{X}')$ is the covariance function for tensorial data with a set of hyperparameters $\boldsymbol{\theta}$. The hyperparameters from observation model and GP prior are collected in $\Theta = \{\sigma, \boldsymbol{\theta}\}$. The model is hierarchically extended to the third level by giving also priors over the hyperparameters in Θ.

To incorporate the knowledge that the training data provides about the function, we use Bayes rule to infer the posterior of the latent function $\mathbf{f} = [f(\mathcal{X}_1), \ldots, f(\mathcal{X}_N)]^T$ by

$$p(\mathbf{f}|\mathcal{D}, \Theta) = \frac{p(\mathbf{y}|\mathbf{f}, \sigma)p(\mathbf{f}|\boldsymbol{\mathcal{X}}, \theta)}{\int p(\mathbf{y}|\mathbf{f}, \sigma)p(\mathbf{f}|\boldsymbol{\mathcal{X}}, \theta)d\mathbf{f}}, \tag{2}$$

where the denominator in (2) is the marginal likelihood obtained by integration over $\mathbf{f}$, yielding

$$\mathbf{y}|\boldsymbol{\mathcal{X}}, \theta, \sigma^2 \sim \mathcal{N}(\mathbf{y} \mid \mathbf{0}, \mathbf{K} + \sigma^2\mathbf{I}), \tag{3}$$

where $(\mathbf{K})_{ij} = k(\mathcal{X}_i, \mathcal{X}_j)$ denotes the covariance matrix or *kernel matrix*. Since the Gaussian observation model is analytically tractable case, which avoids the approximation inference, the conditional posterior of latent function $\mathbf{f}$ is Gaussian, and posterior of f_* is also Gaussian together with the observation y_*. Finally, the predictive distribution of y_* corresponding to $\mathcal{X}_*$ can be inferred as $y_*|\mathcal{X}_*, \mathcal{X}, \mathbf{y}, \Theta \sim \mathcal{N}(\overline{y}_*, \text{cov}(y_*))$, where

$$\begin{aligned} \overline{y}^* &= k(\mathcal{X}_*, \boldsymbol{\mathcal{X}})(k(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}) + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} \\ \text{cov}(y^*) &= k(\mathcal{X}_*, \boldsymbol{\mathcal{X}}_*) - k(\mathcal{X}_*, \boldsymbol{\mathcal{X}})(k(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}}) + \sigma^2\mathbf{I})^{-1}k(\boldsymbol{\mathcal{X}}, \mathcal{X}_*) \end{aligned} \tag{4}$$

3. COVARIANCE FUNCTION WITH TENSOR INPUT

Going beyond a simple vectorial representation of the input data to take into account structure in the input domain is a theme which we see as very important. Although many kernels have been designed for a number of structured objects, few approaches exploit the structure of tensorial representations. Recently, M. Signoretto et. al.[6] proposed a tensorial kernel exploiting algebraic geometry of spaces of tensors and a similarity measure between the different subspaces spanned by higher-order tensors. In addition, they showed that the Hilbert space of multilinear functions associated to general product kernels can be regarded as a space of infinite dimensional tensors. There are some valid reproducing kernels toward a straightforward generalization to Mth-order tensors, such as the kernel functions $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ given as

$$\text{Linear kernel:} \quad k(\mathcal{X}, \mathcal{X}') = \langle \text{vec}(\mathcal{X}), \text{vec}(\mathcal{X}') \rangle,$$

$$\text{Gaussian kernel:} \quad k(\mathcal{X}, \mathcal{X}') = \exp\left(-\frac{1}{2\sigma^2}\|\mathcal{X} - \mathcal{X}'\|^2\right).$$

In order to evaluate the similarity by taking advantage of the multilinear algebraic structure of input tensors, a general product kernel can be defined by M factor kernels, which is valid if the factor kernels are positive semi-definite, denoted by

$$k(\mathcal{X}, \mathcal{X}') = \prod_{m=1}^{M} k\left(\mathbf{X}_{(m)}, \mathbf{X}'_{(m)}\right), \tag{5}$$

where each factor kernel represents a similarity measure between two matrices obtained by mode-m unfolding of two tensor examples. One possibility of similarity measure between matrices is Chordal distance (projection Frobenius norm) on the Grassmannian manifolds. Let $\mathcal{X}$ denote an Mth-order tensor example, SVD can be applied on mode-m unfoldings as $\mathbf{X}_{(m)} = \mathbf{U}_{\mathbf{X}}^{(m)}\mathbf{\Sigma}_{\mathbf{X}}^{(m)}\mathbf{V}_{\mathbf{X}}^{(m)T}$, then the Chordal distance can be computed based on the right singular vectors $\mathbf{V}_{\mathbf{X}}^{(m)}$. As kernels can be interpreted as measures of similarity, it is also possible to define kernels based on information divergences, such as Fisher kernel and *Kullback-Leibler* Kernel [7, 8].

In this study, we applied a probabilistic kernel for tensorial data based on the assumption that each observation in the form of an Mth-order tensor can be individually considered as M different generative models, corresponding to matricization of the tensor in M modes, with a set of observations. For example, we assume $\mathcal{X}_n$ is generated individually by M models governed by parameters $\left\{\boldsymbol{\lambda}_m^{(n)}\right\}_{m=1}^{M}$. Without loss of generality, we apply Gaussian model assumption with the parameters $\boldsymbol{\lambda}$ are priors mean vectors and full covariance matrix. Once the model parameters $\boldsymbol{\lambda}_m$ has been estimated from mode-m matricization $\mathbf{X}_{(m)}$, we can define the kernel distance based on the symmetric *Kullback-Leibler (KL)* divergence, given by

$$\begin{aligned} D\left(p(\mathbf{x}|\boldsymbol{\lambda})||q(\mathbf{x}'|\boldsymbol{\lambda}')\right) &= \int_{-\infty}^{+\infty} p(\mathbf{x}|\boldsymbol{\lambda}) \log\left(\frac{p(\mathbf{x}|\boldsymbol{\lambda})}{q(\mathbf{x}'|\boldsymbol{\lambda}')}\right) d\mathbf{x} \\ &+ \int_{-\infty}^{+\infty} q(\mathbf{x}'|\boldsymbol{\lambda}') \log\left(\frac{q(\mathbf{x}'|\boldsymbol{\lambda}')}{p(\mathbf{x}|\boldsymbol{\lambda})}\right) d\mathbf{x}. \end{aligned} \tag{6}$$

In order to ensure the kernel matrix be a positive definite matrix, we use exponential kernel function based on the symmetric KL divergence measure. Finally, the product kernel from mode-m KL kernels is represented by

$$k(\mathcal{X}, \mathcal{X}') = \prod_{m=1}^{M} \alpha^2 \exp\left(-\frac{D\left(p(\mathbf{X}_{(m)}|\boldsymbol{\lambda}_m)||q(\mathbf{X}'_{(m)}|\boldsymbol{\lambda}'_m)\right)}{2\beta_m^2}\right), \tag{7}$$

where α denots the magnitude hyperparameters and $\beta_1, \ldots, \beta_M$ play the role of characteristic length-scales which implements automatic relevance determination (ARD) [9]. Since the inverse of the length-scale determines how relevant an input is: if the length-scale has a very large value, the covariance will becomes almost independent of that input, effectively removing it from inference. $\boldsymbol{\theta} = \{\alpha, \beta_m | m = 1, \ldots, M\}$ is a vector containing all the hyperparameters of the tensor product kernel.

4. MAP ESTIMATION OF HYPERPARAMETERS

In the tensor-variate covariance function defined in (7), a number of free *hyperparameters* whose values also need to be inferred. In a fully Bayesian approach we should integrate over all unobserved variables. Hence, we can approximate the integral over $p(\boldsymbol{\theta}, \sigma|\mathcal{D})$ with maximum a posterior (MAP) estimate. The marginal likelihood and its partial derivatives w.r.t. the hyperparameters can be obtained by

$$\frac{\partial}{\partial \theta_j} \log p(\mathbf{y}|\boldsymbol{\mathcal{X}}, \boldsymbol{\theta}, \sigma) = \frac{1}{2}\mathbf{y}^T \mathbf{K}_y^{-1} \frac{\partial \mathbf{K}_y}{\partial \theta_j} \mathbf{K}_y^{-1}\mathbf{y} - \frac{1}{2}\text{tr}\left(\mathbf{K}_y^{-1}\frac{\partial \mathbf{K}_y}{\partial \theta_j}\right), \quad (8)$$

where $\mathbf{K}_y = \mathbf{K} + \sigma^2\mathbf{I}$ denotes the covariance matrix for the noisy targets $\mathbf{y}$. The inference on the parameters of covariance functions is conducted mainly transformed space, e.g. log-transformation, which has the advantage that the parameter space is transformed into $(-\infty, +\infty)$.

5. RESULTS

5.1. Simulation on synthetic data

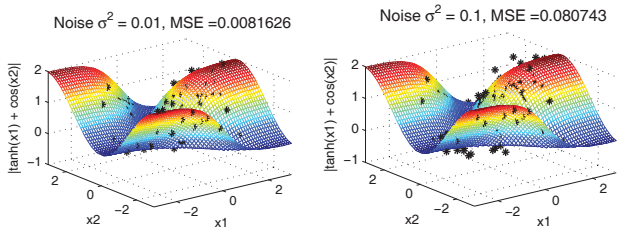

Fig. 1. Estimation of nonlinear function using tensor-based GPs. The surface denotes the actual nonlinear function and balck points denote the predictions of the test set. (a), (b) illustrate the predictions and MSE performance on two datasets generated using different σ^2 .

To demonstrate the effectiveness of tensor-variate based GPs for regression, a set of N random data having 27-dimensions were generated according to $\mathbf{x}_n \sim \mathcal{N}(0, \mathbf{I})$, in order to visualize the results easily, the dependent data was generated using only the first two variate of $\mathbf{x}$ by a nonlinear functions $y_n = |\tanh(x_{n1}) + \cos(x_{n2})| + \varepsilon_n$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The dataset is divided into training and test set. We assume the data $\mathbf{x}_n$ is naturally represented by a higher-order tensor with meaningful modes, thus we reorganized $\mathbf{x}_n$ to $\mathcal{X}_n \in \mathbb{R}^{N\times 3\times 3\times 3}$. The simulation results are illustrated in Fig. 1. Observe that tensor-based GP enables us to predict the nonlinear function accurately without overfitting even though there are many variables in $\mathcal{X}$ that are irrelevant to the dependent data. The MAP estimation of hyperparamters are shown in Table 1. For comparison, we also performed the GPs directly using vectorized data $\mathbf{x}_n \in \mathbf{R}^{N\times 27}$. The results are that $MSE = 0.013$ when $\sigma^2 = 0.01$ and MSE $= 0.087$ when $\sigma^2 = 0.1$. Since the nonlinear function is only related with two variables and the data structure is noninformative for prediction, only slightly improvements using tensor-based GP model are observed. However, it clearly demonstrates the effectiveness of the tensor-based GP model for regression and effectiveness of KL-based kernels for tensorial data. In addition, the number of hyperparameters $\boldsymbol{\theta}$ (i.e., 3) in tensor-based GP is quite smaller than using vectorized GP with 29 hyperparameters, which provides us more robustness to overfitting problem due to the smaller number of parameters.

Table 1. Estimation of hyperparameters

Data sets	$\boldsymbol{\theta}_{\text{MAP}}$	σ^2_{MAP}
$\sigma^2 = 0.01$	(0.49, 1.25, 1.11, 1.42)	0.009
$\sigma^2 = 0.1$	(0.55, 1.07, 1.06, 1.30)	0.01

5.2. Crowd counting from video sequences

There is a great interest in computer vision technology for counting people from video recordings of real environment [10, 11, 12, 13]. In [14], a privacy-preserving system was presented for estimating the count of pedestrians in different directions without using explicit object tracking. However the crowd regions was segmented using a mixture of dynamic textures and various features were extracted from each crowd segment together with a perspective map created manually, resulting in that the successful crowd counting depends on effective crowd segmentation. In this study, higher-order tensor was exploited as structural data representation for each video sequence, and the tensor-variate based GPs can be applied for counting people from videos without explicit segmentation and feature extraction procedure. The pedestrian traffic database was collected from a stationary digital camcorder at University of California, San Diego[1]. Some examples of video are shown in Fig. 2. The original video size is 238×158 with 10 fps, the total 2000 frames were selected for ground-truth annotation and pedestrian count over the region-of-interest (ROI) were marked in every 5 frames of the video. The same setting as [14] was used, i.e. 800 frames for training the model with the remaining 1200 frames for testing, except that each observation is represented by a video sequence containing 5 frames denoted by a tensor $\mathcal{X}_n$, which is consistent with the annotation rate. Fig. 3 illustrates a tensor representation of every 5 frames video sequence and

[1]The detailed description of database can found from http://www.svcl.ucsd.edu/projects/peoplecnt/

mode-m unfolding operation. The mode-m matrices represent M generative models for each observation and is applied for computing the KL divergence based covariance function.

Fig. 2. Examples of scene. Left top image shows the 100th frame with largest crowd bounding box while segmentation of crowd area is shown in left bottom image. Right top image shows a frame with ROI mask, and the normalized complement image is shown in the right bottom.

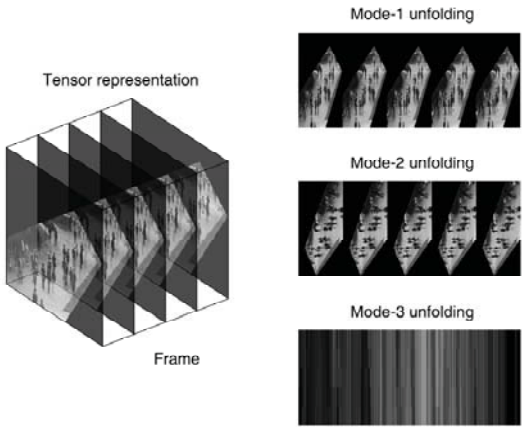

Fig. 3. Each video sequence is represented by a third-order tensor as shown in the left image. The mode-m matricization (unfolding operation) shown in the right are considered as three generative distributions with many corresponding observations, which is used to define the KL divergence based covariance function.

For comparison, we trained the system on two preprocessed datasets: one is the ROI masked videos and the other is further normalized videos by approximately removing the background estimated by mean value along frame mode and complementing the frames as shown in Fig. 2. The prediction performance and MAP estimation of hyperparameters $\boldsymbol{\theta}$ from covariance function and σ^2 from observation model are compared in Table. 2. The results demonstrated the tensor-based GPs successfully predict the crowd counting of pedestrian

Table 2. Crowd counting results using two simplely preprocessed datasets. Mean absolute error and MAP estimation of hyperparameters are shown for comparisons.

	$\lvert\text{error}\rvert$	$\boldsymbol{\theta}_{\text{MAP}}$	σ^2_{MAP}
ROI masked	3.14	(161.4, 12.5, 27.4, 5.5)	0.04
Normalized	3.39	(103.6, 14.8, 29.2, 7.2)	0.03

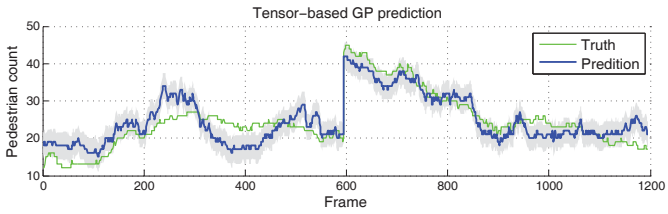

Fig. 4. Crowd counting results over the test sets. The blue line denotes the prediction and green line denotes the ground-truth while gray area shows uncertainty of the predictions.

with averaged absolute error of 3 persons without explicit crowd segmentation and feature extraction. The small difference between two preprocess indicates that tensor-based KL divergence captures the discriminative information automatically thus avoiding the step of removing background. Fig. 4 shows the crowd count estimations for every frame together with uncertainty of predictions. The inappropriate estimation between frames 200-400 is caused by two bicyclists traveling quickly through the scene. Although the performance is just comparable with the state-of-the-art, to our knowledge, this is the first time to perform crowd counting using the raw video data without crowd segmentation and feature extraction.

6. CONCLUSION

We proposed a probabilistic regression framework that brings together the advantages of the GP model and tensor analysis techniques, such that tensor-variate GP regression with Bayesian inference can be performed directly on structured data represented by higher-order tensors. This framework has been shown to allow for a simultaneous account of the multilinear structures and nonlinearity of original data. The advantages of the proposed approach have been demonstrated on a practical example of crowd counting from videos, without the need for explicit segmentation and feature extraction.

Acknowledgments

The work was supported by the JSPS KAKENHI (Grant No. 24700154) and the National Natural Science Foundation of China (Grant Nos. 61202155, 61103122, 91120305, 61272251).

7. REFERENCES

[1] C.E. Rasmussen and CKI Williams, *Gaussian processes for machine learning*, vol. 38, 2006.

[2] J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari, "Bayesian modeling with gaussian processes using the matlab toolbox gpstuff (v3. 3)," *Arxiv preprint arXiv:1206.5754*, 2012.

[3] A. Cichocki, R. Zdunek, A. H. Phan, and S. I. Amari, *Nonnegative Matrix and Tensor Factorizations*, John Wiley & Sons, 2009.

[4] T.G. Kolda and B.W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[5] Z. Xu, F. Yan, and Y. Qi, "Infinite Tucker decomposition: Nonparametric Bayesian models for multiway data analysis," in *Proceedings of the 29 th International Conference on Machine Learning*. 2012.

[6] M. Signoretto, L. De Lathauwer, and J. Suykens, "Kernel-based learning from infinite dimensional 2-way tensors," *Artificial Neural Networks–ICANN 2010*, pp. 59–69, 2010.

[7] P.J. Moreno, P. Ho, and N. Vasconcelos, "A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications," *Advances in Neural Information Processing Systems*, vol. 16, pp. 1385–1393, 2003.

[8] Qibin Zhao, Guoxu Zhou, Tulay Adali, Liqing Zhang, and Andrzej Cichocki, "Kernelization of tensor-based models for multiway data analysis: Processing of multidimensional structured data," *IEEE Signal Processing Magazine*, vol. 30, pp. 137–148, 2013.

[9] D. Wipf, S. Nagarajan, J. Platt, D. Koller, Y. Singer, and S. Roweis, "A new view of automatic relevance determination," *Advances in Neural Information Processing Systems*, vol. 20, pp. 1625–1632, 2008.

[10] Antonio Albiol, Inmaculada Mora, and Valery Naranjo, "Real-time high density people counter using morphological tools," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 2, no. 4, pp. 204–218, 2001.

[11] Yang Cong, Haifeng Gong, Song-Chun Zhu, and Yandong Tang, "Flow mosaicking: Real-time pedestrian counting without scene-specific learning," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1093–1100.

[12] Senem Velipasalar, Ying-Li Tian, and Arun Hampapur, "Automatic counting of interacting people by using a single uncalibrated camera," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 1265–1268.

[13] Vincent Rabaud and Serge Belongie, "Counting crowded moving objects," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 1, pp. 705–711.

[14] A.B. Chan, Z.S.J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE, 2008, pp. 1–7.