Structured Prediction of Music Mood with Twin Gaussian Processes

Santosh Chapaneri(✉) and Deepak Jayaswal

St. Francis Institute of Technology, University of Mumbai, Mumbai, India
{santoshchapaneri,djjayaswal}@sfitengg.org

Abstract. Music mood is one of the most frequently used descriptors when people search for music, but due to its subjective nature, it is difficult to accurately estimate mood. In this work, we propose a structured prediction framework to model the valence and arousal dimensions of mood jointly without requiring multiple regressors. A confidence-interval based estimated consensus from crowdsourced annotations is first learned along with reliabilities of various annotators to serve as the ground truth and is shown to perform better than using the average annotation values. A variational Bayesian approach is used to learn the Gaussian mixture model representation for acoustic features. Using an efficient implementation of Twin Gaussian process for structured regression, the proposed work achieves an improvement in R^2 of 9.3% for arousal and 18.2% for valence relative to state-of-the-art techniques.

Keywords: Music mood · Structured prediction · Crowdsourced annotations

1 Introduction

Mood as a music descriptor is frequently used as a social tag and can be used for organizing large music collection on various smart devices thus requiring modern user interfaces for intuitive music selection and automatic playlist generation. However, the perception of music mood is difficult to quantify due to its subjective nature. Dimensional responses across valence and arousal (VA) are preferred over categorical labels to correspond to the internal human representations as well as to effectively cover all possible mood states [1,2]. It is worthwhile to note that these two dimensions are not completely uncorrelated [3], hence a structured prediction framework [4] is needed to computationally model the affective content of music mood.

1.1 Related Work

The task of music mood estimation is an active research topic [5–8] due to its predominant difficulty in accurately characterizing valence and a comprehensive review of this task is presented in [9]. Music mood being a highly subjective phenomena, multiple annotations per each music clip are required through

B.U. Shankar et al. (Eds.): PReMI 2017, LNCS 10597, pp. 647–654, 2017.
https://doi.org/10.1007/978-3-319-69900-4_82

crowdsourcing to capture the variability of responses. An often over-looked (and less reliable) assumption in most supervised learning techniques for music mood estimation is that the average VA value of multiple annotator responses serves as the ground truth. This leads to the problem of truth discovery analysis for finding the estimated consensus among various annotators [5,10,11]. For feature representation, an appropriate prototype is needed for modeling the acoustic features to avoid the curse of dimensionality [7]. The commonly used regression models such as Support Vector Regression (SVR), Adaboost.RT, and Gaussian Process Regression (GPR) do not handle multi-variate responses, thus 2 different models need to be trained for VA responses. GPR [12] has been shown to outperform SVR for music emotion recognition task [13] due to its capability of hyper-parameter learning and also providing uncertainties in output prediction resulting in soft decision which is suitable for music mood data. Recently, structured regression was performed for computer vision tasks (human pose estimation) [4,14] using Twin Gaussian process (TGP) to predict multi-variate output effectively.

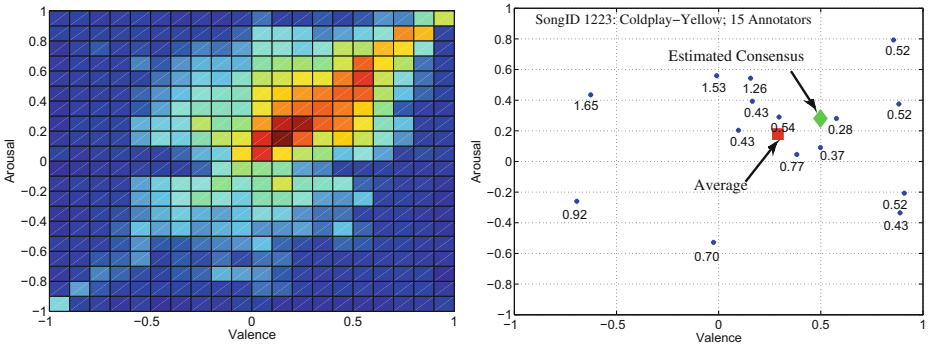
The **contributions** of this work are two-fold: first, we develop an algorithm for determining the estimated consensus of VA values from crowdsourced annotations by tackling the long-tail phenomena; second, we use variational Bayesian inference to compute acoustic posterior features and apply them for structured prediction of music mood using TGP at linear computational cost with a *single* regression model.

2 Proposed Methodology

2.1 Crowdsourced Annotations and Estimated Consensus

We use the AMG1608 dataset [15] that contains crowdsourced annotations of 1608 songs annotated by 665 users, along valence and arousal (VA) dimensions with values in the range $[-1, 1]$. The distribution of all annotations of this dataset is shown in Fig. 1(a), where we observe that the dataset comprises of music clips with mood that cover the complete range with more annotated clips in the 1^{st} quadrant. Since there is no gold standard response available for each music clip, we must estimate it from the available crowdsourced data. A trivial approach is to compute average values along valence and arousal dimensions, however, this assumes that all annotators are equally reliable which may not be true in practice. Extending the work of learning from crowdsourced data [16] for single-dimension target regression case, we derive a maximum-likelihood solution to determine the two-dimensional estimated consensus as well as the reliability (variance) of each annotator.

Consider the dataset $\mathcal{D} = \{\mathbf{y}_i^1, \mathbf{y}_i^2, \ldots, \mathbf{y}_i^R\}_{i=1}^N$ containing two-dimensional VA responses of N music clips by maximum R annotators. Assume a Gaussian distribution model $\mathcal{N}(\mathbf{y}_i^j | \mathbf{y}_i, \alpha^j)$ with α^j as variance of the j^{th} annotator, $\mathbf{y}_i$ as the unknown VA ground truth, and $\mathbf{y}_i^j$ as the response of j^{th} annotator for i^{th} music clip. The parameters to be estimated are $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \mathbf{y}\}$ with the likelihood



(a) (b)

Fig. 1. (a) Annotation distribution of AMG1608 dataset, (b) annotations for a sample clip (*artist*: Coldplay − *song*: Yellow) of AMG1608 with variances of 15 annotators, average and estimated consensus values (Color figure online)

given by Eq. (1) assuming all clips are annotated independently by R annotators. In general, since not all instances will be annotated by each annotator, we define $\mathcal{T}_i$ as the set of annotators providing response for i^{th} clip and $\mathcal{T}_j$ as the set of response provided clips by the j^{th} annotator. Obtaining the gradients of log-likelihood with respect to the parameters results in the maximum likelihood solution given by Eq. (2). This is equivalent to the EM (Expectation-Maximization) algorithm where the E-step determines annotator reliabilities (variances $\hat{\alpha}^j$) and the M-step determines the estimated consensus $\hat{\mathbf{y}}_i$; these two steps are iterated till convergence (e.g. delta change $< 10^{-6}$). The procedure is initialized with $\hat{\mathbf{y}}_i$ as the average of available annotations for i^{th} music clip.

$$P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{j=1}^{R} \frac{1}{\sqrt{2\pi\alpha^j}} \exp\left[\frac{-1}{2\alpha^j}\|\mathbf{y}_i^j - \mathbf{y}_i\|_2^2\right] \tag{1}$$

$$\hat{\alpha}^j = \frac{1}{|\mathcal{T}_j|}\sum_{i\in\mathcal{T}_j}\|\mathbf{y}_i^j - \hat{\mathbf{y}}_i\|_2^2, \quad \hat{\mathbf{y}}_i = \frac{\sum_{j\in\mathcal{T}_i}\mathbf{y}_i^j/\hat{\alpha}^j}{\sum_{j\in\mathcal{T}_i}1/\hat{\alpha}^j} \tag{2}$$

Since the AMG1608 dataset has the (commonly occurring) long-tail problem, the solution of Eq. (2) will be over-optimistic since very few annotators (36 out of 665) provided response to more than 200 music clips and most of the clips were annotated by few users. To handle this problem, we consider the $(1-\beta)$ confidence interval (CI) of annotator reliability where β is the significance value (e.g. 5%). Since the sum of squares of Gaussian random variables follows a Chi-square distribution, we obtain the $(1-\beta)$ confidence interval given by Eq. (3).

$$\frac{1}{\alpha^j}\sum_{i\in\mathcal{T}_j}\|\mathbf{y}_i^j - \hat{\mathbf{y}}_i\|_2^2 = \frac{|\mathcal{T}_j|\hat{\alpha}^j}{\alpha^j} \sim \chi^2(|\mathcal{T}_j|); \quad CI_{1-\beta} = \left\{\frac{|\mathcal{T}_j|\hat{\alpha}^j}{\chi^2_{(1-\frac{\beta}{2},|\mathcal{T}_j|)}}, \frac{|\mathcal{T}_j|\hat{\alpha}^j}{\chi^2_{(\frac{\beta}{2},|\mathcal{T}_j|)}}\right\} \tag{3}$$

$$\hat{\alpha}^j = \frac{1}{\chi^2_{(\frac{\beta}{2},|\mathcal{T}_j|)}} \sum_{i \in \mathcal{T}_j} \|\mathbf{y}_i^j - \hat{\mathbf{y}}_i\|_2^2, \quad \hat{\mathbf{y}}_i = wMedian\left(\mathbf{y}_i^j, \frac{1}{\hat{\alpha}^j}\right)_{j \in \mathcal{T}_i} \tag{4}$$

From Table 1, we observe that annotator ID 542 and ID 647 obtained similar variance $\hat{\alpha}^j$ with Eq. (2), whereas the upper bound (UB) of confidence interval provides a realistic solution. Also, since outliers may exist in the annotated data, instead of removing these outliers with techniques such as minimum covariance determinant (MCD), we propose to use weighted median which is less sensitive to outliers compared to weighted mean. The resulting equations for estimated consensus are given by Eq. (4), which are iterated till convergence. An example of the estimated consensus for a sample clip of AMG1608 is shown in Fig. 1(b) where we observe that annotators having high variance (less reliability) are given less importance for estimating the consensus, whereas the average value gets biased due to few outliers. To further improve the estimated consensus, dependence on input acoustic features needs to be considered; we leave this for future work.

Table 1. Confidence intervals of estimated annotator reliabilities for AMG1608

AnnotatorID	#Annotations	Variance (2)	95% Conf. Int. (3)
159	924	0.4150	(0.3664, 0,4270)
664	240	0.3879	(0.3143, 0,4247)
216	48	0.3724	(0.2774, 0,5462)
542	**12**	**0.4985**	**(0.2709, 1.0901)**
647	**2**	**0.4981**	**(0.1903, 11.1160)**

2.2 Bayesian Acoustic GMM Feature Representation

For each music clip in the dataset, we compute standard acoustic features across four categories: dynamics (root-mean-square energy), spectral (centroid, spread, skewness, kurtosis, entropy, flatness, 85% roll-off, 95% roll-off, brightness, roughness, irregularity), timbral (zero-crossing rate, flux, 13-dimensional MFCCs, delta MFCCs, delta-delta MFCCs) and tonal (key clarity, musical mode, harmonic change likelihood, 12-bin chroma vector, chroma peak, chroma centroid), resulting in 70-dimensional feature vector per frame of 50 msec duration with 50% overlap [7]. Each feature dimension is normalized to zero mean and unit standard deviation. Block-level features are used to capture temporal characteristics across frames where each block comprises of 16 consecutive frames with overlap of 12 frames. The block-level feature vector $\mathbf{r}_t$ consists of mean and standard deviation of frame-based feature vectors [8]. To represent the block-level features as a fixed dimensional vector, [13] computes the mean and standard deviation across all frames and stacks these into a single vector. For an effective prototypical representation, we adopt the EM-GMM clustering approach of [7]

resulting in AGMM (Acoustic Gaussian Mixture Model) posterior probabilities x_{nk} given by Eq. (5), where the universal background model (UBM) parameters $\{\pi_k, \mu_k, \Sigma_k\}$ indicate the weight, mean and covariance of k^{th} latent audio topic (or mixture), and $\mathbf{x}_n$ is the $1 \times K$ acoustic posterior probability feature vector of song s_n consisting of F_n blocks. Using EM algorithm [17], the UBM model is trained with randomly selected 25% block-level feature vectors across the entire dataset (spanning whole range of VA space) resulting in $215,000$ vectors.

$$p(\mathbf{r}_t) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{r}_t|\mu_k, \Sigma_k)\,;\; x_{nk} = \frac{1}{F_n} \sum_{t=1}^{F_n} \left(\frac{\pi_k \mathcal{N}(\mathbf{r}_t|\mu_k, \Sigma_k)}{\sum_{h=1}^{K} \pi_h \mathcal{N}(\mathbf{r}_t|\mu_h, \Sigma_h)} \right) \tag{5}$$

The most crucial problem of AGMM is determining the exact number of latent audio topics K that can explain the given data. In [7], various values of K (16, 32, 64, 128, 256, 512) were used to determine the regression performance. However, this is ad-hoc as it could lead to over-fitting of data. To determine the optimal number of mixtures, we resort to variational Bayesian inference framework [17] resulting in Bayesian Acoustic GMM (BAGMM) posterior probability feature representation for each music clip. With the Bayesian treatment, all parameters of AGMM are given priors: the weight is assigned Dirichlet prior and mean and covariance are assigned Gaussian-Wishart prior. We refer the reader to [17] (Chap. 10) for detailed formulation of variational posteriors and variational EM algorithm. The algorithm is initialized with α_0 (hyper-parameter of Dirichlet prior) as 0.001, $\mathbf{m}_0$ (hyper-parameter of Gaussian prior) as k-means centroid of training data (to speed up the convergence), and $\mathbf{W}_0$ (hyper-parameter of Wishart prior) as $10\mathbf{I}$ to avoid mixtures getting trapped in local maximum (here, $\mathbf{I}$ is the identity matrix). The expected value of mixing weights in the posterior distribution of BAGMM is given by Eq. (6) with N_k as the responsibility of k^{th} mixture, where we observe that for uninformative priors ($\alpha_0 \to \infty$), the expected value converges to a small value of ξ ($0 < \xi < 1/K$).

$$E[\pi_k] = \frac{\alpha_k + N_k}{K\alpha_0 + \sum_{j=1}^{K} N_j} = \frac{\alpha_0 + 2N_k}{K\alpha_0 + N} \approx \frac{\alpha_0/N}{K\alpha_0/N + 1} = \frac{1}{K + N/\alpha_0} \tag{6}$$

BAGMM determines the optimal number of latent audio topics without using cross-validation ($K_{opt} = 117$ in this work), solves the problem of singularity [17] that occurs in AGMM, and prevents over-fitting of data. Each music clip is thus represented as $1 \times K_{opt}$ acoustic posterior probability feature vector $\mathbf{x}_n$, with the corresponding estimated consensus $\hat{\mathbf{y}}_n$ (cf. Sect. 2.1), denoted as $\mathbf{y}_n$ hereafter.

2.3 Structured Prediction

The conventional regression approach of building independently two different regressors does not consider any correlation that may exist in the multi-variate response. From the estimated consensus of AMG1608 dataset, we find that the correlation coefficient between valence and arousal is 0.312, which should be accounted for during training of the regression model. We follow the work of [4]

that introduced Twin Gaussian Processes (TGP) with Kullback-Leibler (KL) divergence measure for structured prediction to model not only the input but also the output covariance for effective regression performance. A generalized version of TGP was proposed in [14] with Sharma-Mittal (SM) divergence measure and was shown to perform better than KLTGP for structured prediction. However, both KLTGP and SMTGP suffer from high computational complexity since they need to solve a non-linear optimization problem requiring L-BFGS solver. An efficient approach termed as direct TGP (dTGP) was proposed in [18] that require only $O(N)$ computations compared to $O(N^2)$ of KLTGP and SMTGP.

$$l_{\mathbf{y}}^{'} = \min\left(\left(1+\lambda_y\right)\mathbf{1}, \max\left(\mathbf{0}, \hat{\mu} K_Y u_{\mathbf{x}}\right)\right);\ u_{\mathbf{x}} = K_X^{-1} K_X^{\mathbf{x}};\ \eta = K_X - {K_X^{\mathbf{x}}}^T u_{\mathbf{x}} \quad (7)$$

$$\hat{\mu} = \frac{-\eta + \sqrt{\eta^2 + 4u_{\mathbf{x}}^T K_Y u_{\mathbf{x}}\left(1+\lambda_y\right)}}{2u_{\mathbf{x}}^T K_Y u_{\mathbf{x}}};\ \hat{\gamma}_m = \frac{l_{\mathbf{y},m}^{'}}{\sum_{m=1}^{M} l_{\mathbf{y},m}^{'}};\ \mathbf{y}^{'} = \sum_{m=1}^{M} \hat{\gamma}_m \mathbf{y}_m^{'} \quad (8)$$

Given N training instances of input BAGMM feature vectors $\mathbf{X} \in \mathbb{R}^{N \times K_{opt}}$ and corresponding output estimated consensus $\mathbf{Y} \in \mathbb{R}^{N \times 2}$, dTGP optimizes the value of output kernel function $l_{\mathbf{y}} = K_Y^{\mathbf{y}}$ with its solution $l_{\mathbf{y}}^{'}$ shown in Eq. (7) obtained via simple algebra, instead of optimizing for response variable $\mathbf{y}$ directly as is done in KLTGP and SMTGP. To simplify computations, dTGP further finds M nearest neighbors for a specific test input based on $\hat{\gamma}_m$ and computes the predicted bi-variate output $\mathbf{y}^{'}$ as the weighted sum of M nearest training instances ($M = 30$ set empirically) given by Eq. (8). The weight $\hat{\gamma}_m$ captures the input-output as well as between-output correlation for structured prediction [18]. The input and output kernel functions for Gaussian processes are given by Eq. (9) with the kernel parameters (ρ = kernel bandwidth, λ = regularization parameter) determined experimentally via a grid-search [4].

$$K_X\left(\mathbf{x}_i, \mathbf{x}_j\right) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\rho_x^2}} + \lambda_x \delta_{ij};\ K_Y\left(\mathbf{y}_i, \mathbf{y}_j\right) = e^{-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2\rho_y^2}} + \lambda_y \delta_{ij} \quad (9)$$

3 Experimental Results

We perform evaluation on AMG1608 dataset with two metrics: R^2 (coefficient of determination) and $RMSE$ (root mean square error) to measure the regression performance between predicted VA values and the estimated consensus.

Table 2 shows the performance of existing state-of-the-art techniques and proposed work. For [13], conventional acoustic features aggregated with mean and standard deviation were used and GPR was used with combination of Squared Exponential and Rational Quadratic kernels [12]. For histogram density modeling (HDM) of mood annotations, the number of latent histograms was set to $K = 256$ as reported in [8]. For [6], conventional acoustic features were used with adaptive aggregation of GP regressors. For GPR [13] and Aggregate GPR [6], two independent regression models were used for valence and arousal estimation with average annotation value set as the ground truth, while HDM [8] needed fine-tuning of number of latent histograms to model each latent audio

Table 2. Performance of estimating music mood with 10-fold cross-validation

Method	R^2_{Arousal}	$RMSE_{\text{Arousal}}$	R^2_{Valence}	$RMSE_{\text{Valence}}$
GPR [13]	0.654 $\pm$0.058	0.247 $\pm$0.011	0.429 $\pm$0.082	0.283 $\pm$0.073
HDM ($G = 7$) [8]	0.632 $\pm$0.045	0.258 $\pm$0.013	0.352 $\pm$0.056	0.291 $\pm$0.027
Aggregate GPR [6]	0.678 $\pm$0.095	0.203 $\pm$0.010	0.437 $\pm$0.053	0.236 $\pm$0.013
KLTGP [4]	0.717 $\pm$0.061	0.184 $\pm$0.015	0.502 $\pm$0.049	0.237 $\pm$0.046
SMTGP [14]	**0.721** $\pm$0.026	**0.165** $\pm$0.055	**0.512** $\pm$0.028	**0.224** $\pm$0.014
Proposed	**0.715** $\pm$0.022	**0.172** $\pm$0.043	**0.507** $\pm$0.021	**0.235** $\pm$0.037

topic as well as choice of grid size $G \times G$ to represent the VA space as a heatmap. We evaluate the structured prediction performance of TGPs using the proposed BAGMM feature representation and estimated consensus set as ground truth with a single regression model. Though SMTGP performs slightly better than KLTGP, it incurs quadratic computational complexity due to inversion of kernel matrices during optimization. The performance of proposed work with dTGP is similar to KLTGP but at a linear computational cost. All values were calculated with 10 different combinations of training and test data and the means and standard deviations of these values are reported to measure the performance (10-fold cross-validation). Overall, we observe an improvement in R^2 of 9.3% for arousal and 18.2% for valence relative to GPR [13].

4 Conclusion

A structured prediction framework is presented in this work for affective modeling of music mood using TGPs and is shown to perform better than independently learning different regressors for valence and arousal dimensions. An EM-type algorithm is proposed to determine the confidence-interval based estimated consensus to serve as the ground truth for regression. Using Bayesian inference, acoustic features are represented with BAGMM posterior probabilities where the optimal number of Gaussian mixtures are determined automatically from the training data and the drawback of singularity and over-fitting of conventional GMM is simultaneously avoided. The proposed framework achieves improvement in music mood prediction with direct TGP implementation that achieves optimization with simple algebra and avoids the use of non-linear quadratic optimization of KLTGP and SMTGP. The future work includes deriving estimated consensus by considering dependence of response on input acoustic features and solving the problem of covariate shift [18] across various music mood datasets.

Acknowledgments. The authors thank Yi-Hsuan Yang for sharing the music clips of AMG1608 dataset and Mohamed Elhoseiny for sharing SMTGP implementation via personal communication.

References

1. Hu, X.: Music and mood: Where theory and reality meet. In: Proceedings of the 5th iConference, Chicago, USA (2010)
2. Brinker, B., Dinther, R., Skowronek, J.: Expressed music mood classification compared with valence and arousal ratings. EURASIP J. Audio, Speech Music Process. **24**, 1–14 (2012)
3. Kumar, N., Guha, T., Huang, C., Vaz, C., Narayanan, S.: Novel affective features for multiscale prediction of emotion in music. In: Proceedings of the 18th IEEE International Workshop on Multimedia Signal Processing (MMSP), Montreal, Canada (2016)
4. Bo, L., Sminchisescu, C.: Twin Gaussian processes for structured prediction. Springer Int. J. Comput. Vis. **87**(28), 1–25 (2010)
5. Chin, Y., Wang, J., Wang, J., Yang, Y.: Predicting the probability density function of music emotion using emotion space mapping. IEEE Trans. Affect. Comput. **PP**(99), 1–10 (2016)
6. Fukayama, S., Goto, M.: Music emotion recognition with adaptive aggregation of Gaussian process regressors. In: Proceedings of the 41st IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China (2016)
7. Wang, J., Yang, Y., Wang, H., Jeng, S.: Modeling the affective content of music with a Gaussian mixture model. IEEE Trans. Affect. Comput. **6**(1), 56–68 (2015)
8. Wang, J., Wang, H., Lanckriet, G.: A histogram density modeling approach to music emotion recognition. In: Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia (2015)
9. Yang, Y., Chen, H.: Machine recognition of music emotion: a review. ACM Trans. Intell. Syst. Technol. **3**(3), 1–30 (2012)
10. Wan, M., Chen, X., Kaplan, L., Han, J., Gao, J., Zhao, B.: From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California (2016)
11. Ramakrishna, A., Gupta, R., Grossman, R., Narayanan, S.: An expectation maximization approach to joint modeling of multidimensional ratings derived from multiple annotators. In: INTERSPEECH, San Francisco, USA (2016)
12. Rasmussen, C., Williams, C.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
13. Markov, K., Matsui, T.: Music genre and emotion recognition using Gaussian processes. IEEE Access **2**, 688–697 (2014)
14. Elhoseiny, M., Elgammal, A.: Generalized twin Gaussian processes using Sharma-Mittal divergence. Springer J. Mach. Learn. **100**(2), 399–424 (2015)
15. Chen, Y., Yang, Y., Wang, J., Chen, H.: The AMG1608 dataset for music emotion recognition. In: Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia (2015)
16. Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., Moy, L.: Learning from crowds. J. Mach. Learn. Res. **11**, 1297–1322 (2010)
17. Bishop, C.: Pattern Recognition and Machine Learning. Springer, New York (2006)
18. Yamada, M., Sigal, L., Chang, Y.: Domain adaptation for structured regression. Int. J. Comput. Vis. **109**(2), 126–145 (2014)