TWO-LEVEL MULTI-TASK METRIC LEARNING WITH APPLICATION TO MULTI-CLASSIFICATION

Hong Liu, Xuewu Zhang, Pingping Wu

Key Laboratory of Machine Perception (Ministry of Education)
Engineering Lab on Intelligent Perception for Internet of Things (ELIP)
Shenzhen Graduate School, Peking University, China
{hongliu,xuewuzhang,pingpingwu}@pku.edu.cn

ABSTRACT

Many metric learning approaches neglect that the real world multi-class problems share strong visual similarities, which can be exploited by learning discriminative models. In this paper, a Two-level Multi-task Metric Learning (TMTL) method is presented to learn a distance measure from equivalence constraints. Multiple features are adopted to represent the image information and learn the distance matrices in the first level. Then the task-specific learning paradigm and multi-task voting mechanism make full use of pairwise equivalence labels, which induces knowledge from anonymous pairs to multi-classification. Experiments are conducted on two challenging benchmarks PubFig and OuluVS for face identification and lipreading respectively. The results demonstrate that our method outperforms the recent multi-task learning approaches and multi-class support vector machine.

Index Terms— Metric Learning, Multi-task Learning, Face Identification, Lipreading

1. INTRODUCTION

Metric learning is an emerging field aiming at a more powerful and discriminative distance from labeled examples. Recently there has been considerable interest for distance metric learning (DML) with various applications. Particularly, it can significantly improve the performance for face recognition [1, 2], person re-identification [3], image retrieval [4] or tracking [5]. Also, a comprehensive survey can be found in [6].

Most current works on metric learning focus on the Mahalanobis distance learning [1, 2, 7, 8]. Specifically, the KISS Metric Learning (KISSME) [9] is designed to deal with general pairwise constraints, which does not rely on a tedious iterative optimization and performs efficiently. Although remarkable successes have been achieved on many small scale problems, they cannot be directly transplanted to large scale visual applications with the following disadvantages.

Firstly, many existing metric learning methods learn Mahalanobis distance metric from a single feature for each image. However, how to learn a similarity measure with multiple features has rarely been discussed in DML, since it cannot deal with multiple feature representations directly. Specifically, the KISSME and other Mahalanobis distance learning methods usually learn one metric without considering how to jointly learn multiple matrices. Besides, the simple feature concatenation is in a risk of over-fitting and computation complexity growth, which is not appropriate.

Secondly, it usually involves hundreds of classes for real world applications. However, for most multi-classification problems, we find that images from the same class subset usually share some common properties [10], while images from different subsets intend to be very easy to be classified. As with this problem, several metric learning approaches have been proposed. Parameswaran *et al.* [11] extended the well performed LMNN [2] to multi-task metric learning, and Grauman *et al.* [12] proposed to learn a tree of metrics to incorporate the object hierarchy. In contrast, multi-task learning (MTL) [13] approaches a cluster of similar tasks in parallel in order to improve the performance on all tasks.

Accordingly, we mainly concern about the two aspects mentioned above. Inspired by Multiple Kernel Learning [14], we propose to learn the Mahalanobis matrices from multiple features and define a unified distance metric. Basically, it is desirable to learn distance matrices from these multiple features so that more discriminative information can be exploited, which we call the first level. Further, some problems such as face recognition or lipreading share strong visual similarities and person-specific dissimilarities, which can be exploited from learning discriminative models. Therefore, we extend the efficient KISSME algorithm to the multi-task paradigm, which induces knowledge from anonymous pairs to multi classification.

This work is supported by National Natural Science Foundation of China (NSFC, No.61340046, 60875050, 60675025), Science and Technology Innovation Commission of Shenzhen Municipality (No.JCYJ20120614152234873, No.JCYJ20130331144631730, No.JCYJ20130331144716089), Specialized Research Fund for the Doctoral Program of Higher Education (No.20130001110011).

2. ALGORITHM DESCRIPTION

Our Two-level Multi-task Metric Learning (TMTL) method takes both the multiple features and specific task into consideration. As a result, it can handle large scale data and multi-class problems. To introduce our approach, we give an overview of our framework. Then, we introduce our TMTL approach and the multi-task voting scheme, which allows to a multi-class decision by using multiple metrics.

2.1. Framework Overview

To demonstrate our two-level multi-task metric learning explicitly, face identification, which can be regarded as a multi-classification problem, is taken as an example to illustrate. As Fig. 1 shows, to identify a face, a specific metric is learned for each person by using multiple features, which is regarded as a subtask in the first level. Then a joint metric is learned from all subtasks. To derive a metric, pairwise examples are employed. Particularly, a similar pair indicates two face images from the same person defined as positive samples, while a dissimilar pair is from the specific person and the rest.

After the training procedure, both the specific metric and the joint metric are obtained. For a test image, the distances are computed between the image and all positive images using the learned metrics. Then, a Multi-Task voting scheme is adopted and decide which specific class(subtask) it belongs to, which will be detailed described as follows.

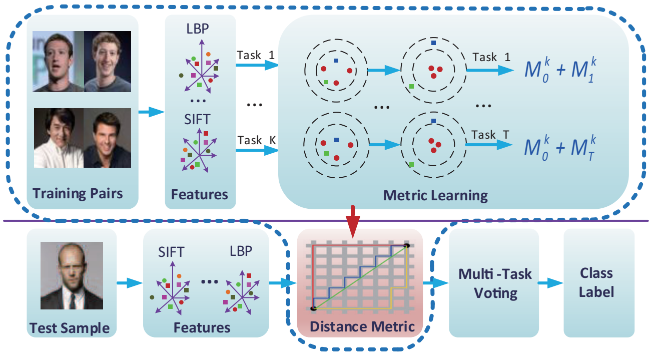

Fig. 1. Illustration of the proposed approach for face identification.

2.2. Two-level Multi-Task Metric Learning

The Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j) \tag{1}$$

which measures the squared distance between two data points $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$. The learning task is to induce a valid (pseudo) metric and make sure $\mathbf{M}$ is a symmetric positive semi-define matrix. The KISSME [9] learns a Mahalanobis metric motivated by a statistical inference perspective based on a likelihood-ratio test, where the Mahalanobis distance matrix $\mathbf{M}$ is obtained by

$$\mathbf{M} = \Sigma_S^{-1} - \Sigma_D^{-1} \tag{2}$$

Algorithm 1: Two-level Multi-Task Metric Learning

```
Input: Training pairs $\{\mathbf{x}_i^k, \mathbf{x}_j^k\} \in \mathbb{R}^{d_k}, k = 1, 2, ..., K$,
       $K$ is the number of feature types. Parallel task
       $t \in \{1, 2, ..., T\}$.
Output: Projection matrices $\hat{\mathbf{M}}_t^k \in R^{(d_k \times d_k)_t}$.
Define a task-specific subset of similar pairs $S_t$ and
dissimilar $D_t$.
Compute weight of task specific balance factor $\mu$.
for $k \leftarrow 1$ to $K$ do
    for $t \leftarrow 1$ to $T$ do
        Extract $k$th feature for $S_t$ and $D_t$.
        Calculate task-specific matrix of $k$th feature as
        shown in Eq. (4).
        Compute shared metric in Eq. (5).
        Obtain $\hat{\mathbf{M}}_t^k$ in Eq. (6).
    end
end
return $\hat{\mathbf{M}}_t^k \in R^{(d_k \times d_k)_t}$.
```

Based on KISSME, we extend it to Two-level Multi-Task Learning (TMTL) paradigm. In the first level, multiple features are considered, while in the second level, we model the information sharing mechanism among different learning tasks. Finally, the overall distance can be calculated by the learned multiple matrices.

Formally, denoting that $k \in \{1, 2, ..., K\}$ is the feature type in the first level, for each feature we model the individual metric for each task $t \in \{1, 2, ..., T\}$ as a combination of a shared metric $\mathbf{M}_0^k$ and a task-specific metric $\mathbf{M}_t^k$. Given a pair of training samples, we define the distance as:

$$d_t^2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{K} (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{M}_0^k + \mathbf{M}_t^k)(\mathbf{x}_i - \mathbf{x}_j) \tag{3}$$

Each task defines a task-specific subset of similar and dissimilar samples pairs: $S_t = \{(i, j) \in \psi_t | y_i = y_j\}$ and $D_t = \{(i, j) \in \psi_t | y_i \neq y_j\}$. According to Eq. (2), the task-specific matrix of kth feature can be calculated by

$$\mathbf{M}_t^k = \mathbf{M}_{S_t}^{-1} - \mathbf{M}_{D_t}^{-1}$$
$$= \left(\frac{1}{|S_t|} \sum_{(i,j)\in S_t} \mathbf{C}_{ij} \right)^{-1} - \left(\frac{1}{|D_t|} \sum_{(i,j)\in D_t} \mathbf{C}_{ij} \right)^{-1} \tag{4}$$

where $\mathbf{C}_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)$. Thus, the common metric can be estimated by averaging all individual tasks, which can be formulated as:

$$\mathbf{M}_0^k = \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{M}_{S_t} \right)^{-1} - \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{M}_{D_t} \right)^{-1} \tag{5}$$

Then, the final individual Mahalanobis distance metric is given by

$$\hat{\mathbf{M}}_t^k = \mathbf{M}_0^k + \mu \mathbf{M}_t^k \tag{6}$$

Correspondingly, $\mathbf{M}_0^k$ models commonalities across all tasks for each feature. In contrast, $\mathbf{M}_t^k$ focuses on task-specific characteristics, which has a balancing factor μ between the task specific metric $\mathbf{M}_t^k$ and the shared metric $\mathbf{M}_0^k$. Hereby, μ is the proportion of subtask samples in total.

To make our learning procedures clear, the procedures are listed in Algorithm 1. Due to the efficiency and convenience of original KISS metric learning, our framework works without much optimization and other time cost operations. Besides, the multi-task mechanism can also handle multi-class problems thank to the task-specific characteristic.

2.3. Multi-Task Voting

To make full use of our two-level multi-task metric learning method, all task-specific metrics are combined into a multi-class decision. Similar to the majority voting scheme, the strategy [15] is adopted which is to assign the class that wins most pairwise comparisons.

Assume that the positive samples for task t are coincidence with class label $\mathbf{x}_i : y_i = t$. Given a test sample $\mathbf{x}_i$, the voting scheme is adopted to decide which task label the sample belongs to. In order to improve the robustness of comparison, not only the individual distance metric of task t is compared, but also the complementary distance metric of task $u \in \{u \neq t | u = 1, ...T\}$. Then the final decision is made for class that gets most pairwise comparisons. Details are shown in Algorithm 2.

3. EXPERIMENTS AND DISCUSSIONS

To show the applicability of our method we conduct experiments on two different standard benchmarks with rather diverse characteristics. The goals of our experiments are two folds. First, we want to show that the TMTL can be applied for face identification with anonymous pairwise labels. Accordingly, we compare our results to standard metric learning and related multi-task learning approaches. Second, we want to show that our method can handle lipreading which is a typical multi-class problem [16–18]. Due to the task-specific characteristic, our methods induces knowledge from anonymous pairs to multi-classification. Therefore, we evaluate the performance compared to the multi-class SVMs.

3.1. Face Identification

Face recognition is a general topic that includes both face identification (who is it) and face verification (deciding if two faces match). Generally speaking, it is easy to have the equivalence labels, but face identification need class labels for individual identity. Additionally, for face identification it is not obvious how to make use of this anonymous information.

Algorithm 2: Multi-Task Voting Scheme

```
Input: Distance matrices $\hat{\mathbf{M}}_t^k \in R^{(d_k \times d_k)_t}$. Index set
       $\psi_t$ of each specific task. Test sample $\mathbf{x}_i$.
Output: Class label $c \in \{1, 2, ..., T\}$.
Initialize:
    $I_t \leftarrow 0, Label(t) \leftarrow 0, c \leftarrow 0, A \leftarrow 0, B \leftarrow 0.$
for $t \leftarrow 1$ to $T$ do
    while $u \neq t$, and $u \in \{1, 2, ..., T\}$ do
        Calculate $A \leftarrow \min_{j \in \psi_t \wedge y_j = t} d_t^2(\mathbf{x}_i, \mathbf{x}_j)$.
        if $A \leq \min_{k \in \psi_u \wedge y_k = u} d_t^2(\mathbf{x}_i, \mathbf{x}_k)$ then
            $I_t \leftarrow I_t + 1.$
        end
        Calculate $B \leftarrow \min_{j \in \psi_t \wedge y_j = t} d_u^2(\mathbf{x}_i, \mathbf{x}_j)$.
        if $B \leq \min_{k \in \psi_u \wedge y_k = u} d_u^2(\mathbf{x}_i, \mathbf{x}_k)$ then
            $I_t \leftarrow I_t + 1.$
        end
    end
    $Label(t) \leftarrow I_t$.
end
$c \leftarrow \arg\max_t = Label(t)$.
return Class label c.
```

Here, we want to show that the TMTL can take advantage of class labels for identification problem.

In the following, we demonstrate the performance of our method on the Public Figures Face Database (PubFig) [19]. The number of images of per individual ranges from 63 to 1536. Some illustrative examples are given in Figure 2. Similar to the existing verification protocol in 10 folds for cross-validation, we split the images in 10 non-overlapping folds for cross-validation in the evaluation set which contains 42461 images of 140 individuals.

Fig. 2. Examples of evaluation set in PubFig database [19].

To extract image features, three descriptors including LBPs [20], SIFT [1] and the "high-level" description of visual face traits [19] are used. Then we demonstrate the performance by comparing it to recent MTL methods [11] and also benchmark to multi-class support vector machines [21]. Similar to [22], we rank the classifier scores and compare with different thresholds. Thus, we obtain the recall-precision curve where the recall means the percentage of samples having higher score than the current threshold while the precision means the ratio of correctly classified samples.

In Figure 3 (a), we compare TMTL to the single-task KISSME and multi-task KISSME as described in [22]. The TMTL method outperforms both two methods. With TMTL we reach the accuracy of 71.20% at full recall. Moreover, the multi-task is proved better than the single-task.

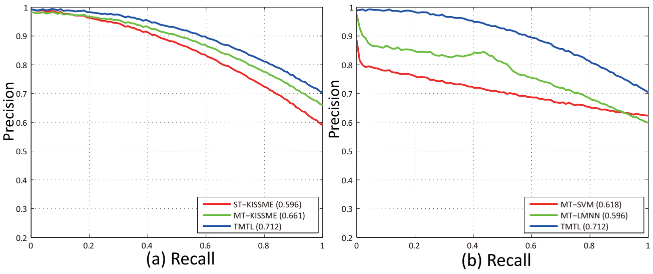

Fig. 3. Comparison of TMTL to (a) single-task learning and MT-KISSME in [22], (b) other MTL and MT-SVM.

Next, we benchmark to recent MTL methods MT-LMNN [11] and MT-SVM [21]. In order to keep the consistency of comparison, all the methods are implemented using the same features as TMTL does. As shown Figure 3 (b), the standard multi-class one-vs-all SVM reaches with 61.80% while MT-LMNN reaches with 59.6% at full recall. However, the TMTL beats both by 9.4% and 10.6%. Admittedly, the TMTL gains information from pairwise labels and the task-specific learning is favorable.

3.2. LipReading

The OuluVS dataset [16] consists of 20 subjects uttering 10 phrase five times at resolution of 720 × 576. The specific task is to learn a distance metric for each phrase with pairwise equivalence labels. However, there is no predefined set or procedure to obtain dissimilar pairs. Hence, we generate dissimilar pairs by randomly combining utterance of different phrases, while the similar pairs are from the same phrase.

To compare our method to other approaches, we did the same preprocessing and experimental setting described in [17]. Specifically, for the Subject Independent (SID) experiment, the leave-one-subject-out is explored in the uttering database. In particular, each phrase corresponds a sub-task and there are 19 × 5 utterances with total C_{95}^2 similar pairs, which are all from the same phrase. While the dissimilar pairs are composed by combining each current phrase utterance and the rest different phrase utterances which are randomly selected. For the Subject Dependent (SD) experiment, the leave-one-utterance-out is performed. That is, training sets are the first 4 utterances of each phrase from all speakers. There are 20 × 4 utterances which has total C_{80}^2 similar pairs. Similarly, the dissimilar pairs are generated by randomly selecting the rest different phrase utterances.

To represent the utterance, three descriptors LBPs [16], HOG [23] and MIP described in [24] are adopted. Then all the feature vectors are projected into a low dimensional subspace by PCA and the training is performed in all phrase-specific task parallelly. Further, the multi-task voting scheme is adopted to decide which phrase the test utterance belongs to. After the same procedure is repeated on each speaker (SID) and each utterance (SD) respectively, the overall result is obtained. To compare with multi-SVMs, experiments are conducted with the same feature in [16, 23]. In Fig. 4, we compare the relative performance change for per phrase with single-task learning, multi-SVMs and TMTL with single feature. Accordingly, most phrases can be improved by TMTL. Finally, the results of Lipreading are listed in Table I. Our method outperforms the one-vs-all SVMs in [16, 17]. Further, multiple features in TMTL produce an improvement to less features. Also, TMTL can match the state-of-the-arts RFMA [23]. Essentially, it is attributed to both learning the more discriminative models from multiple features and inducing anonymous information in a multi-task way. Besides, the sharing mechanism among different learning tasks also contribute to getting information from pairwise labels.

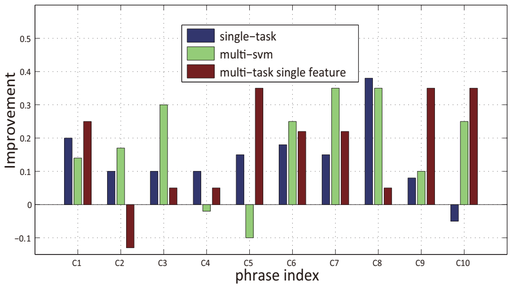

Fig. 4. Accuracy improvement on per phrase

Table I. Lipreading results of SID and SD experiments on OuluVS.

Methods	Classifier	Accuracy (%)	
		SID	SD
LBP-TOP [16]	SVM	58.85	64.20
Curve Match [17]	SVM	81.30	85.10
RFMA$_{HOG+LBP}$ in [23]	Distance Matching	86.40	93.60
RFMA$_{fusion}$ [23]	Distance Matching	**89.70**	**97.30**
TMTL$_{LBP-TOP}$	Multi-Task Voting	85.45	86.57
TMTL$_{HOG+LBP}$	Multi-Task Voting	87.25	92.76
TMTL$_{multi-feature}$	Multi-Task Voting	**88.94**	**95.82**

4. CONCLUSIONS

We propose a Two-level Multi-task Metric Learning (TMTL) method to learn a distance metric from equivalence constraints in a multi-level view. Our method jointly learns multiple matrices from multiple features in a multi-task paradigm. It exploits the shared similarities as well as task-specific information among different learning tasks, which is scalable to large datasets and capable to multi-class problems. Essentially, our model allows knowledge transfer from anonymous pairs which combines the information both effective image understanding and multi-task learning. Moreover, experiment results also shows that our method has a wide range of applications, which can be exploited in the future.

5. REFERENCES

[1] M. Guillaumin, J. Verbeek, and C. Schmid, “Is that you? metric learning approaches for face identification,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 498–505, Sept 2009.

[2] K. Q. Weinberger and L. K. Saul, “Distance metric learning for large margin nearest neighbor classification,” *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[3] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, “Pedestrian recognition with a learned metric,” in *Asian Conference on Computer Vision (ACCV)*, pp. 501–512, 2011.

[4] S. C. Hoi, W. Liu, M. R. Lyu, and W. Y. Ma, “Learning distance metrics with contextual constraints for image retrieval,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 2072–2078, 2006.

[5] X. Wang, G. Hua, and T. Han, “Discriminative tracking by metric learning,” in *European Conference on Computer Vision (ECCV)*, pp. 200–214, 2010.

[6] B. Kulis, “Metric learning: A survey,” *Foundations & Trends in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2012.

[7] A. Mignon and F. Jurie, “Pcca: A new approach for distance learning from sparse pairwise constraints,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2666–2672, June 2012.

[8] V. D. Jason, K. Brian, J. Prateek, S. Suvrit, and I. S. D., “Information-theoretic metric learning,” in *International Conference on Machine Learning (ICML)*, pp. 209–216, June 2007.

[9] M. Köstinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof, “Large scale metric learning from equivalence constraints,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2288–2295, June 2012.

[10] A. Torralba, K. P. Murphy, and W. T. Freeman, “Sharing visual features for multiclass and multiview object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 29, no. 5, pp. 854–869, 2007.

[11] S. Parameswaran and K. Q. Weinberger, “Large margin multi-task metric learning,” in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1867–1875, 2010.

[12] K. Grauman, F. Sha, and S. Hwang, “Learning a tree of metrics with disjoint visual features,” in *Advances in Neural Information Processing Systems (NIPS)*, pp. 621–629, 2011.

[13] R. Caruana, “Multitask learning,” *Machine Learning Journal (MLJ)*, vol. 28, pp. 41–75, July 1997.

[14] F. Bach, G. Lanckriet, and M. Jordan, “Multiple kernel learning, conic duality, and the smo algorithm,” in *International Conference on Machine Learning (ICML)*, p. 6, 2004.

[15] J. Friedman, “Another approach to polychotomous classification,” *Technical report, Department of Statistics, Stanford University*, 1996.

[16] G. Zhao, M. Barnard, and M. Pietikainen, “Lipreading with local spatiotemporal descriptors,” *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[17] Z. Zhou, G. Zhao, and M. Pietikainen, “Towards a practical lipreading system,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 137–144, 2011.

[18] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, “Extraction of visual features for lipreading,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002.

[19] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, “Attribute and Simile Classifiers for Face Verification,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 365–372, Oct 2009.

[20] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[21] C. Chang and C. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[22] M. Köstinger, P. Roth, and H. Bischof, *Synergy-based learning of facial identity*. Springer, 2012.

[23] Y. Pei, T. Kim, and H. Zha, “Unsupervised random forest manifold alignment for lipreading,” in *IEEE International Conference on Computer Vision (ICCV)*, pp. 129–136, 2013.

[24] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, “Motion interchange patterns for action recognition in unconstrained videos,” in *European Conference on Computer Vision (ECCV)*, pp. 256–269, Springer, 2012.