CONTRIBUTION-BASED FEATURE TRANSFER FOR JPEG MISMATCHED STEGANALYSIS

Chaoyu Feng, XiangWei Kong, Ming Li, Yong Yang, Yanqing Guo

School of Information and Communication Engineering
Dalian University of Technology, Dalian, Liaoning, 116024, China
Email:{fengchaoyu, yongyang}@mail.dlut.edu.cn, {kongxw, mli, guoyq}@dlut.edu.cn

ABSTRACT

In realistic steganalysis applications, the mismatched problem can lead to the degradation of performance in steganalysis. The main reason is the discrepancy of feature distributions between training set and testing set. In this paper, we present a Contribution-based Feature Transfer (CFT) algorithm for JPEG mismatched steganalysis. CFT tries to learn two transformations to transfer training set features by evaluating both the sample feature and dimensional feature contributions. We can obtain new feature representations so as to approach the feature distribution of the testing samples. The comparison to prior arts reveals the superiority of CFT on the experiments for the mismatched JPEG steganalysis in the heterogeneous cover source scenario.

Index Terms— Mismatched steganalysis, feature transfer, contribution, JPEG image.

1. INTRODUCTION

Steganography is the science of hiding data into the public digital media. Many universal steganographic approaches for JPEG image have been proposed, such as MME [1], F5 [2], MBS [3], OutGuess [4] and nsF5 [5]. Diversity strategies, such as statistics-preserving, heuristic algorithms and minimal distortion, have been used to increase the undetectability of hidden data in media [5]. In contrast, steganalysis aims to identify the existence of the hidden data in the given media. More and more steganalysis algorithms have achieved satisfactory detection performance (the detection accuracy above 90%) even in the low embedding rate [6]-[7].

However, the achievement is based on the assumption that training and testing sets are sampled from identical feature distribution. In steganalysis, the differences of feature distributions between training and testing samples will result in the phenomenon called mismatch. The mismatched steganalysis can lead to the degradation (the detection accuracy almost below 75%) of detection accuracy [8]-[9]. Facing this serious problem, Ker and other famous professors in information security area have appealed to improve the practicability of steganalysis in the real world [10]. More and more researchers have devoted the study of different paradigms of mismatched steganalysis [11].

In [12], Ker *et al.* proposed mishmash strategies for different classifiers to mitigate the model mismatch. Fridrich *et al.* [9] designed two kinds of algorithms, Mixture and Closest, for cover source mismatch by comparing the gap of quantification table between training and testing samples. In [8], Lubenko *et al.* used the large and diverse data to train simple classifiers for solving the mismatched problem. Xu *et al.* [13] constructed large representative training set to reduce the intra-class variation for cover source mismatch. In [14], Pibre *et al.* used Convolutional Neural Network framework with big filters in convolution layer and improved the detection for mismatched problem. However, these methods need to collect sufficient and diverse labeled samples. It is very tough to train steganalysis classifiers with many steganographic algorithms.

To avoid the tedious re-collection of various training data, one solution is to migrate knowledge from training data to testing data for learning better steganalyzers. In [15]-[16], Kong *et al.* learnt shared feature space by introducing transfer learning and improved detection performance of mismatched steganaysis. In [17], Zeng *et al.* proposed two complementary criterions on the basis of transfer learning and relieved the influence of quantization table mismatch.

Inspired by these studies, we notice that improving similarity of feature distribution can be valid for the mismatched problem. Unlike [16]-[17], however, we consider that many features in the training set can enlarge the differences of feature distribution and be useless for getting a better steganalyzer. This motivated us to choose useful features in training set to learn better feature representations. In this paper, we consider mismatched steganalysis in heterogeneous cover source scenario, which is an universal phenomenon, and propose a novel Contribution-based Feature Transfer (CFT) approach. CFT evaluates appropriate feature contribution for training set to construct effective training features by using two transfers for both sample feature and dimensional feature, which can make the discrepancy of feature distribution between training samples and testing samples smaller.

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

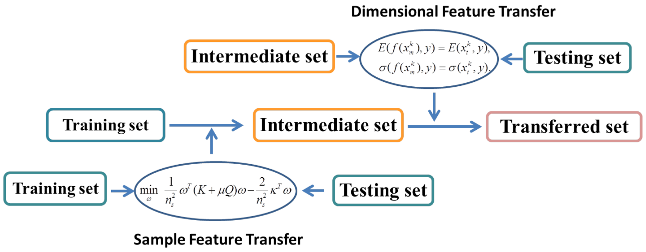

Fig. 1. Illustration of contribution-based feature transfer.

The remainder of this paper is organized as follows: proposed Contribution-based Feature Transfer algorithm is presented in Section 2. In Section 3, we conduct experiments and compare our approach with previous arts. Finally, we summarize this paper in Section 4.

2. PROPOSED METHOD

In the CFT, we firstly transfer the training set to the intermediate set by finding the appropriate contribution values for each training sample. This so-called Contribution-based Sample Feature Transfer can reduce the distribution discrepancy between training and testing samples in the Reproducing Kernel Hilbert Space and keep little inter-class correlation. Then we learn a transformation, which can distribute more contribution to dimensional features that occur frequently in the testing set, to make intra-class distribution in the intermediate set close to that in the testing set. This step is referred as Contribution-based Dimensional Feature Transfer. From two views, sample and dimension, we reduce the discrepancy of feature distributions between training and testing set sufficiently and get new feature representations for learning a better steganalyzer. The procedure of contribution-based feature transfer is illustrated in Fig.1. The details of our method will be described as follows.

2.1. Contribution-based Sample Feature Transfer

Suppose we have a set n_s of training feature samples $X_s = \{x_{s1}, x_{s2}, \ldots, x_{sn_s}\} \in \mathbb{R}^d$, where d is the dimension of feature vector. Let $y_{si} \in \{0,1\}$ be the label of x_{si}, $i = 1, 2, \ldots, n_s$, and $\{0,1\}$ presents cover image and stego image, respectively. Similarly, we have a set of n_t testing feature samples $X_t = \{x_{t1}, x_{t2}, \ldots, x_{tn_t}\} \in \mathbb{R}^d$, a set of unknown labels $y_{tj} \in \{0,1\}$ of x_{tj}, $j = 1, 2, \ldots, n_t$.

We utilize a non-parametric method called Kernel Mean Matching (KMM) [18] to estimate the differences of feature distribution in the Reproducing Kernel Hilbert Space (RKHS). Let $\omega \in \mathbb{R}^{n_s}$ be suitable values of contribution. KMM computes the distribution discrepancy between the means of two sample sets. The objective function is to minimize:

$$KMM(X_s, X_t) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \omega_i \phi(x_{s_i}) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi(x_{tj}) \right\|_H^2, \tag{1}$$

where $\phi(\cdot)$ is kernel-induced feature mapping to the RKHS and $\| \cdot \|_H$ denotes the ℓ_2 norm in the RKHS, ω_i is the value of contribution for x_{s_i}.

However, the formula (1) have just considered the relationship between training and testing samples. It ignores the relationship between samples and labels. We adopt correlation coefficient to keep inter-class relationship of sample feature smaller according to their corresponding labels. So we design new correlation coefficient formula as:

$$\min_{\omega} \frac{1}{n_s^2} \omega^T Q \omega, \tag{2}$$

where $\frac{1}{n_s^2}$ restricts the importance of each correlation coefficient. $Q \in \mathbb{R}^{n_s \times n_s}$ describe the relationship between different features. The element of Q can be computed as:

$$Q_{ij} = \begin{cases} 0, & \text{if } y_{s_i} = y_{s_j}, \\ \rho_{ij}, & \text{otherwise }, \end{cases} \tag{3}$$

$$\rho_{ij} = \frac{\sum_{k=1}^{d} (x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{d} (x_{ik} - \overline{x}_i)^2 \sum_{k=1}^{d} (x_{jk} - \overline{x}_j)^2}}, \tag{4}$$

where x_{ik} represents the kth dimension feature of the ith sample. $\overline{x}_i$ is the mean of the ith sample feature.

In order to ensure ω rational, we add two constraints: $\omega_i \in [LB, UB]$ and $|\frac{1}{n_s} \sum_{i=1}^{n_s} \omega_i - 1| \leq \epsilon$. Combining (1) and (2), from the view of sample, we not only minimize the differences of distribution between two sets, but also decrease the inter-class correlation for training samples. The final optimization formula can be represented as:

$$\begin{aligned} \min_{\omega} \quad & \frac{1}{n_s^2} \omega^T (K + \mu Q) \omega - \frac{2}{n_s^2} \kappa \omega, \\ \text{s.t.} \quad & |\frac{1}{n_s} \sum_{i=1}^{n_s} \omega_i - 1| \leq \epsilon, \\ & LB \leq \omega_i \leq UB, \end{aligned} \tag{5}$$

where $K_{ij} := k(x_{s_i}, x_{s_j}) = \phi(x_{s_i})^T \phi(x_{s_j})$ and $\kappa_i := \frac{n_s}{n_t} \sum_{j=1}^{n_t} k(x_{s_i}, x_{t_j})$, μ is trade-off parameter to adjust the effect of corresponding term. The formula (5) can be regarded as a quadratic programming (QP) by finding suitable ω and be solved by interior point methods or any other optimization algorithms. When ω is available, we can transfer samples in the training set to the intermediate set $X_m = \{x_{m_1}, x_{m_2}, \ldots, x_{m_{n_s}}\}$, where x_{m_i}represents intermediate sample which is transformed from the sample x_{s_i}. Let $y_{mi} = y_{si}$ be the label of x_{mi}.

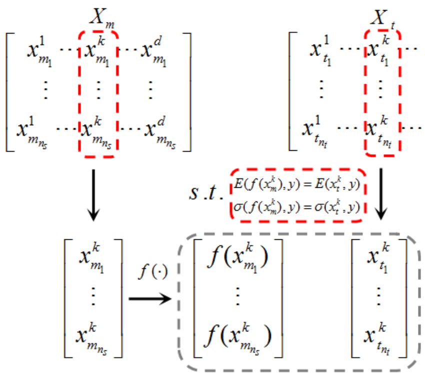

Fig. 2. Illustration of contribution-based dimensional feature transfer.

2.2. Contribution-based Dimensional Feature Transfer

Besides the view of sample feature, we further increase the feature distribution similarity from the perspective of dimension. We notice that matching statistic, which can reflect the characteristic of feature distribution, can be helpful to match the feature distribution between two sets. We give more contribution for dimensional features that occur frequently in the testing set to increase the feature statistic similarity of the same category by applying a class-based transformation to the intermediate set. The procedure is illustrated in Fig.2

Let x_m^k be the kth dimension feature of intermediate samples, and x_t^k be the kth dimension feature of testing samples, $k = 1, 2, \ldots, d$. Let $E(x_m^k, y)$ and $\sigma(x_m^k, y)$ represent the joint expectation and standard deviation of the kth dimension feature of samples with label y in the intermediate set, where $y \in \{0, 1\}$. Similarly, $E(x_t^k, y)$ and $\sigma(x_t^k, y)$ represent those in the testing set.

We try to match the statistic (joint expectation and standard deviation) of each dimension feature between the intermediate set and testing set in each category. We propose a liner transformation $f(\cdot)$ for each dimension of samples in the intermediate set to reduce the intra-class statistic discrepancy:

$$E(f(x_m^k), y) = E(x_t^k, y), \sigma(f(x_m^k), y) = \sigma(x_t^k, y). \quad (6)$$

To reach the goal in (6), we follow the maximum entropy phraseology and $f(\cdot)$ can be obtained by

$$f(x_{m_i}^k) = \left(x_{m_i}^k - E(x_m^k, y = y_{mi})\right) \frac{\sigma(x_t^k, y = y_{mi})}{\sigma(x_m^k, y = y_{mi})} + E(x_t^k, y = y_{mi}). \quad (7)$$

However, since the labels are unknown in the testing set, $E(x_t^k, y)$ and $\sigma(x_t^k, y)$ cannot be computed directly. To solve the problem, we adopt a simple approach which is similar to [19] to find the estimate $\hat{E}(x_t^k, y)$ and $\hat{\sigma}(x_t^k, y)$. We use a universal classifier (such as support vector machine) to train on the intermediate samples and obtain the posterior probability estimate $\hat{p}_t(y|x_{t_j})$ on the unlabeled testing samples. Hence, the approximation $\hat{E}(x_t^k, y)$ and $\hat{\sigma}(x_t^k, y)$ can be computed by

$$\hat{E}(x_t^k, y) \approx \frac{1}{\sum_{j=1}^{n_t} \hat{p}_t(y|x_{tj})} \sum_{j=1}^{n_t} x_{t_j}^k \hat{p}_t(y|x_{tj}), \quad (8)$$

$$\hat{\sigma}_{(}x_t^k, y) \approx \sqrt{\frac{1}{\sum_{j=1}^{n_t} \hat{p}_t(y|x_{tj})} \sum_{j=1}^{n_t} (x_{t_j}^k - \hat{E}(x_t^k, y))^2 \hat{p}_t(y|x_{tj})}. \quad (9)$$

After getting $\hat{E}(x_t^k, y)$ and $\hat{\sigma}(x_t^k, y)$, we can use (7) to obtain the transferred sample features and transferred set.

2.3. Iterative Optimization

As we can see that our transfer process includes two parts. ω in (5) and $\hat{p}_t(y|x_{t_j})$ in (8) (9) can both influence the performance of feature contribution measuring. However, these can be stable and reliable when we iterate the two steps above by using the transferred sample features. So we propose to optimize ω and $\hat{p}_t(y|x_{t_j})$ by running two steps above iteratively. We minimize the differences of both ω and $\hat{p}_t(y|x_{t_j})$ between twice of iterations. The formula is to minimize:

$$\sum_{j=1}^{n_t} \sum_{y_{tj} \in y} |\hat{p}_t^r(y|x_{tj}) - \hat{p}_t^{r-1}(y|x_{tj})| + \sum_{i=1}^{n_s} |\omega_i^r - \omega_i^{r-1}|, \quad (10)$$

where r is the r-time of iterations, $r = 2, 3, \ldots, IT$. IT is the maximum iteration time. After several iterations, (10) becomes stable and converged. We can obtain the ultimate transferred feature representations. Traditional machine learning can be used to train classification or regression models on these features. Our proposed method is summarized in Algorithm 1.

Algorithm 1 Contribution-based Feature Transfer

Input:
Training set X_s, testing set X_t, training set labels y_s
Output: Classifier h; values of contribution ω

Transform training set X_s to an intermediate set X_m by (5) and get values of contribution ω;
Train classifier h on intermediate set and get the posterior probability $\hat{p}_t(y|x_{t_j})$ for testing samples;
repeat
    Transfer intermediate set X_m to the transferred set by (7);
    Train classifier h on the transferred set and update posterior probability for testing set X_t;
    Transform the transferred set to an intermediate set X_m by (5) and update the values of contribution ω;
until (10) is converged
return Classifier h and values of contribution ω.

Table 1. Detection accuracy of JPEG quality factor 75.

Train-Test	Mixture	TCA	KMM	IMFA	CFT
F5-MBS	0.655	0.690	0.708	0.748	**0.763**
F5-nsF5	0.630	0.590	0.615	0.643	**0.675**
MBS-F5	0.577	0.598	0.728	0.773	**0.828**
MBS-nsF5	0.527	0.552	0.585	0.593	**0.615**
nsF5-F5	0.802	0.778	0.740	0.830	**0.835**
nsF5-MBS	0.690	0.673	0.652	0.802	**0.815**
Average	0.647	0.647	0.671	0.731	**0.755**

3. EXPERIMENTS

In this section, we evaluate our proposed CFT approach on steganographic algorithm mismatched condition for JPEG steganalysis in the heterogeneous cover source scenario. The heterogeneous cover source often appears in real life and the steganographic algorithm is a common mismatched factor which can demonstrate the effective performance of our proposed algorithm.

3.1. Experimental settings

Based on the widely use of JPEG images, we carry out experiments in JPEG domain. The public dataset we used is the BOSSbase 1.01 database [20], which includes 10,000 8-bits gray scale images of size 512×512. The cover images are created by compressing the original database with standard quantization tables for JPEG quality factors 75, 85 and 95. With different JPEG quality factors, we can obtain different cover sources.

To create stego images, three steganographic algorithms (F5 [2], MBS [3], nsF5 [5]) are chosen to embed the message into cover images. These three steganographic algorithms are common and widespread use in JPEG domain. The payload is set to 10% of the maximum embedding capacity. With different steganographic algorithms, we have three sets and each set includes two categories, cover images and stego images. We randomly select 300 labeled images per category for each JPEG quality factor in one steganographic algorithm to construct training set with 1800 images in heterogeneous cover source. Similarly, we select another 300 unlabeled images per category for each quality factor in another steganographic algorithm as testing set.

Considering that using JRM features [7] will lead to a high computational complexity, we choose the 274-dimensional PEV features [6] as the original features in our experiments. We use lib-SVM [21] [22] as a classifier. We compare our proposed CFT algorithm with several state-of-the-art methods, including Mixture [9], TCA [23], KMM [18] and IMFA [16].

Table 2. Detection accuracy of JPEG quality factor 85.

Train-Test	Mixture	TCA	KMM	IMFA	CFT
F5-MBS	0.632	0.682	0.707	0.737	**0.752**
F5-nsF5	0.617	0.625	0.647	0.677	**0.707**
MBS-F5	0.577	0.595	0.715	0.760	**0.817**
MBS-nsF5	0.512	0.545	0.595	0.616	**0.640**
nsF5-F5	0.795	0.787	0.737	0.818	**0.822**
nsF5-MBS	0.740	0.710	0.683	0.827	**0.835**
Average	0.648	0.657	0.681	0.739	**0.762**

Table 3. Detection accuracy of JPEG quality factor 95.

Train-Test	Mixture	TCA	KMM	IMFA	CFT
F5-MBS	0.725	0.715	0.790	0.840	**0.870**
F5-nsF5	0.690	0.660	0.697	0.682	**0.735**
MBS-F5	0.625	0.622	0.722	0.800	**0.830**
MBS-nsF5	0.535	0.555	0.623	0.643	**0.672**
nsF5-F5	0.795	0.795	0.777	0.832	**0.840**
nsF5-MBS	0.770	0.740	0.720	0.840	**0.847**
Average	0.690	0.681	0.722	0.773	**0.799**

3.2. Experimental Results

In our experiments, we repeat 5 times for each mismatched situation. The performance is evaluated by the maximum equal-prior accuracy rate p_{avg} which is calculated by $p_{avg} = \frac{1}{2}max(p_c + p_s)$. p_c and p_s represent the cover classification accuracy rate and the stego classification accuracy rate, respectively. The maximum is taken over parallel decision boundaries [12]. The results are shown in Tables 1, 2 and 3. From the experimental results, we can find that our proposed algorithm outperforms the other state-of-the-art methods. The results demonstrate that feature transfer based on the appropriate contribution for training samples can obtain more discriminate feature representations.

4. CONCLUSION

This paper proposed a novel mismatched steganalysis algorithm Contribution-based Feature Transfer. Our method derived discriminate features from original features in training set. We improved the distribution similarity between the training set and testing set by transferring both the sample and dimensional feature which measured the contribution of training set feature. Experimental results illustrated the effectiveness of our method for JPEG mismatched steganalysis in heterogeneous cover source scenario.

5. ACKNOWLEDGEMENTS

This work is supported by the Foundation for Innovative Research Groups of the NSFC (Grant No. 71421001) and NSFC (Grant No. 61172109).

6. REFERENCES

[1] Y. Kim, Z. Duric, and D. Richards, "Modified matrix encoding technique for minimal distortion steganography," in *Proc. Int. Workshop Information Hiding*, Alexandria, Virginia, July 2006, pp. 314–327.

[2] A. Westfeld, "F5-a steganographic algorithm: High capacity despite better steganalysis," in *Proc. Int. Workshop Information Hiding*, Pittsburgh, Pennsylvania, Apr. 2001, pp. 289–302.

[3] P. Sallee, "Model-based steganography," in *Proc. Int. Workshop Digital Watermarking*, Seoul, SouthKorea, Oct. 2004, pp. 154–167.

[4] N. Provos, "Defending Against Statistical Steganalysis," *Usenix security symposium*, vol. 10, pp. 323–336, Aug. 2001.

[5] J. Fridrich, T. Pevnỳ, and J. Kodovskỳ, "Statistically undetectable jpeg steganography: dead ends challenges, and opportunities," in *Proc. ACM Multimedia and Security Workshop (MMSec)*, Dallas, Texas, Sep. 2007, pp. 3–14.

[6] T. Pevnỳ and J. Fridrich, "Merging markov and DCT features for multiclass jpeg steganalysis," in *Proc. SPIE Security, Steganography, and Watermarking of Multimedia Contents*, San Jose, California, Feb. 2007, pp. 650503–650503.

[7] J. Fridrich and J. Kodovskỳ, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, May, 2012.

[8] I. Lubenko and A. D. Ker, "Steganalysis with mismatched covers: Do simple classifiers help?," in *Proc. ACM Multimedia and Security Workshop (MMSec)*, Coventry, UK, Sep. 2012, pp. 11–18.

[9] J. Kodovskỳ, V. Sedighi, and J. Fridrich, "Study of cover source mismatch in steganalysis and ways to mitigate its impact," in *Proc. SPIE Media Watermarking, Security, and Forens.*, San Francisco, California, Feb. 2014, pp. 90280J–90280J.

[10] A. D. Ker, P. Bas, R. Böhme, R. Cogranne, S. Craver, T, Filler, J. Fridrich, and T. Pevnỳ, "Moving steganography and steganalysis from the laboratory into the real world," in *Proc. ACM Information Hiding and Multimedia Security Workshop*, Montpellier, France, June 2013, pp. 45–58.

[11] X. Zhao, J. Zhu, and H. Yu, "On More Paradigms of Steganalysis," *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 8, no. 2, pp. 1–15, Apr. 2016.

[12] A. D. Ker and T. Pevnỳ, "A mishmash of methods for mitigating the model mismatch mess," in *Proc. SPIE Media Watermarking, Security, and Forens.*, San Francisco, California, Feb. 2014, pp. 90280I–90280I.

[13] X. Xu, J. Dong, W. Wang, and T. Tan, "Robust steganalysis based on training set construction and ensemble classifiers weighting," in *Proc. IEEE Int. Conf. Image Process (ICIP)*, Quebec City, Canada, Sep. 2015, pp. 1498–1502.

[14] L. Pibre, P. Jérôme, D. Ienco, and M. Chaumont, "Deep learning is a good steganalysis tool when embedding key is reused for different images, even if there is a cover sourcemismatch," in *SPIE Electronic Imaging*, Feb. 2016.

[15] X. Li, X. Kong, B. Wang, Y. Guo, and X. You, "Generalized transfer component analysis for mismatched JPEG steganalysis," in *Proc. IEEE Int. Conf. Image Process (ICIP)*, Melbourne, Australia, Sep. 2013, pp. 4432–4436.

[16] X. Kong, C. Feng, M. Li, and Y. Guo, "Iterative multi-order feature alignment for JPEG mismatched steganalysis," *Neurocomputing*, vol. 214, pp. 458–470, Jun. 2016.

[17] L. Zeng, X. Kong, M. Li, and Y. Guo, "JPEG quantization table mismatched steganalysis via robust discriminative feature transformation," in *Proc. SPIE Media Watermarking, Security, and Forens.*, San Francisco, California, Feb. 2015, pp. 94090U–94090U.

[18] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," *Dataset shift in machine learning*, vol. 3, pp. 5, 2009.

[19] N. Farajidavar, T. deCampos and J. Kittler, "Adaptive transductive transfer machines," in *Proc. British Machine Vision Conference (BMVC)*, Nottingham, UK, Sep. 2014.

[20] T. Filler, T. Pevnỳ, and P. Bas, "BOSS, (Break Our Steganography System)," *http://www.agents.cz/boss/*.

[21] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, May 2011.

[22] P. Bas and T. Furon, "Bows-2," *http://bows2.gipsa-lab.inpg.fr*.

[23] S. Pan, I. Tsang, J. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.