Toward Intelligent Product Retrieval for TV-to-Online (T2O) Application: A Transfer Metric Learning Approach

Qiang Fu, Yong Luo, Yonggang Wen, *Senior Member, IEEE*, Dacheng Tao, *Fellow, IEEE*, Ying Li, and Ling-Yu Duan, *Member, IEEE*

***Abstract*—It is desired (especially for young people) to shop for the same or similar products shown in the multimedia contents (such as online TV programs). This indicates an urgent demand for improving the experience of TV-to-Online (T2O). In this paper, a transfer learning approach as well as a prototype system for effortless T2O experience is developed. In the system, a key component is high-precision product search, which is to fulfill exact matching between a query item and the database ones. The matching performance primarily relies on distance estimation, but the data characteristics cannot be well modeled and exploited by a simple Euclidean distance. This motivates us to introduce distance metric learning (DML) for improving the distance estimation. However, in traditional DML methods, the side information (such as the similar/dissimilar constraints or relevance/irrelevance judgements) in the target domain is leveraged. These methods may fail due to limited side information. Fortunately, this issue can be alleviated by utilizing transfer metric learning (TML) to exploit information from other related domains. In this paper, a novel manifold regularized heterogeneous multitask metric learning framework is proposed, in which each domain is treated equally. The proposed approach allows us to simultaneously exploit the information from other domains and the unlabeled information. Furthermore, the ranking-based loss is adopted to make our model more appropriate for search. Experiments on two challenging real-world datasets demonstrate the effectiveness of the proposed method. This TML approach is expected to impact the transformation of the emerging T2O trend in both TV and online video domains.**

***Index Terms*—TV-to-Online, distance metric learning, transfer learning, heterogeneous domains, manifold regularization, ranking-based loss.**

I. Introduction

THE way that multimedia contents (such as photographs and videos) are consumed has been transformed by the current era of Mobile Internet due to the growing popularity of the smart mobile devices (e.g., smartphone and laptop). Specifically, the experience of consuming video contents in the main screen (e.g., TV) and having access to the companion contents in a second device (e.g., smartphone and tablet) has become widely appreciated for viewers. Such a multi-screen video experience [1] has in turn led to an emerging business model, TV-to-Online (T2O). It bridges the gap between video contents and online merchants. With the help of T2O systems, the video viewers are able to quickly locate the desired products, which is same or similar with the item displayed on video program. This impulse purchase can be done online through the second screen. For example, many items from the Korea drama "My Love from the Star" are very attractive for worldwide consumers. When watching this drama and enthralled by some lipstick from a particular scene, the T2O users can easily complete the purchase from online merchants via their mobile phones. This activity has been touted as an innovative model that enables us to fulfill the "I want" moment as pointed by Google [2].

Inspired by this emerging market trend and based on the highly touted multi-screen social TV system [3], we developed an effortless T2O subsystem. In our system, people are allowed to buy the desired items via online merchants. Particularly, this purchase behavior can be done simultaneously while watching video contents on the web or TV. The T2O system is composed of several modules, and product search (or retrieval) is among the most significant ones. After capturing the desired product from the video scene, the search function aims to match the queried item with the online merchant list. Two major

Manuscript received August 10, 2017; revised November 28, 2017; accepted December 16, 2017. Date of publication January 10, 2018; date of current version July 17, 2018. This work was supported in part by Singapore NRF2015ENC-GDCR01001-003 (administrated via IMDA), NRF2015ENC-GBICRD001-012 (administrated via BCA), by Tier 1 projects (RG17/14, RG26/16), by the Data Science and Artificial Intelligence Research Centre, Nanyang Technological University, and by Australian Research Council Projects FL-170100117, DP-180103424, DP-140102164, and LP-150100671. This work was also supported by the Key Research and Development Program of Beijing Municipal Science and Technology Commission (No. D171100003517002), in part by the National Natural Science Foundation of China under Grant U1611461 and Grant 61661146005, and by the National Key Research and Development Program of China under Grant 2016YFB1001501. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Honggang Wang. *(Corresponding author: Ling-Yu Duan.)*

Q. Fu and Y. Li are with the School of Software and Microelectronics, Peking University, Beijing 100080, China (e-mail: qiang.fu@pku.edu.cn; li.ying@pku.edu.cn).

Y. Luo and Y. Wen are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: yluo180@gmail.com; ygwen@ntu.edu.sg).

D. Tao is with the UBTECH Sydney Artificial Intelligence Centre and the School of Information Technologies, Faculty of Engineering and Information Technologies, University of Sydney, Darlington, NSW 2008, Australia (e-mail: dacheng.tao@sydney.edu.au).

L.-Y. Duan is with the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing 100080, China (e-mail: lingyu@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2018.2791803

stages are involved in the search (or retrieval) problem [4], [5]: 1) both the features of the query item and reference items (such as products) in the database (e.g., merchant list) are extracted; 2) the similarity or distance of each pair of items (the query item versus each item listed in the database) are calculated [6]. Hence, an appropriate distance estimation strategy plays a critical role in achieving satisfactory performance.

To improve the user experience of the T2O system, an appropriate distance estimation algorithm is required. Euclidian (EU) metric based distance estimation and the direct utilization of original features may fail in such application, since both the important information of the target problem (product search) and statistical properties among the features are ignored. To overcome this issue, distance metric learning (DML) [7], [8] is thus introduced. The side information (similar/dissimilar constraints or relevance/irrelevance judgements) contained in the target problem can be fully exploited by DML to learn a feature-specific metric. In this paper, we introduce DML to the T2O system for distance estimation. To learn a reliable metric by DML, good amount of side information is needed. However, it is usually insufficient in real-world applications since the labeling cost is high. In this scenario, DML may fail. This problem can be alleviated by transfer metric learning (TML) [9], which is able to leverage the knowledge (such as side information) from other related domains [10]. In our application of product search, multiple modalities of the data may exist. For example, the images of certain products may be associated with some text descriptions and hyperlinks. Moreover, a variety of visual features [11], [12] can be extracted to represent an image.

In this paper, each feature space or modality is regarded as a domain, and a novel manifold regularized heterogeneous MTML (MRHMTML) framework is developed for improving the product search in our T2O system by effectively utilizing the side information from each domain. We also assume there are abundant multi-domain unlabeled samples, each of them has representations in all domains. Specifically, metrics of all different domains are learned in a single optimization problem, where the empirical loss w.r.t. each domain is minimized. Meanwhile, the metric learning is reformulated as learning feature transformation [13]. We project the different representations of the given unlabeled samples into a common subspace and maximize their high-order [14], [15] covariance in the subspace. This results in improved feature transformations since the side information of all domains are utilized to learn the shared subspace. Intuitively, the common subspace bridges different domains so that information can be successfully transferred. The learned metrics are thus more reliable than learning them separately. This is particularly beneficial when the side information is limited. Moreover, a manifold regularization term [16] is added to make full use of the unlabeled information in each domain by exploring the geometric structure of the data.

Our algorithm is superior to other related methods. For example, transformations of multiple heterogeneous domains are also learned together in [17] and [18]. However, these approaches only explore the statistics (correlation information) between pairs of representations in either one-vs-one [18], or centralized [17] way. Thus, the high-order statistics are ignored, which can only be obtained by examining all domains simultaneously. Our approach outperforms them in that:

1) More information is utilized to learn the metrics since the high-order correlations of all domains are exploited, which may contribute to better performance;
2) The unlabeled data are well exploited by enabling knowledge transfer across domain and preserving topology in each domain;
3) The ranking based loss is adopted to learn metrics, which elegantly supports product search.

Extensive experiments are conducted on the product subsets of two challenging social image datasets: PASCAL VOC [19] and NUS WIDE [20]. We compare our method with not only Euclidean (EU) and single domain ranking-based DML baselines [6], [13], but also a representative heterogeneous multi-task learning approach [18]. Effectiveness of the proposed RHMTML is demonstrated by the promising results. For example, we have an on average more than 10% relative improvements compared with the EU baseline on the PASCAL VOC dataset in terms of MAP.

II. Related Work

Our work is mainly related to distance metric learning and heterogeneous transfer learning.

A. Distance Metric Learning

The goal of distance metric learning (DML) is to learn an appropriate distance function over the input space, so that the relationships between data are appropriately reflected. Most conventional metric learning methods, which are often called "Mahalanobis metric learning", can be regarded as learning a linear transformation of the input data [21], [22]. The first work of Mahalanobis metric learning was done by Xing *et al.* [7], where a constrained convex optimization problem with no regularization was proposed. Some other representative algorithms include the neighborhood component analysis (NCA) [23], large margin nearest neighbors (LMNN) [24], information theoretic metric learning (ITML) [25], etc.

These algorithms are developed for clustering and classification. To learn metric for information retrieval, some ranking based metric learning approaches have been proposed [6], [13], [26]. In [6], the authors indicate that the "must-link" and "cannot-link" constraints used in the traditional DML are suboptimal for information retrieval. A ranking based loss is designed to address this problem by separating distances between query and relevant samples from distances between query and irrelevant samples. Ranking SVM was extended to learn distance metric in [26], and a scalable DML algorithm that optimizes ranking measure via stochastic gradient descent (SGD) is proposed in [13] to handle large datasets.

Recently, transfer metric learning (TML) has attracted intensive attention to tackle the labeled data deficiency issue in the target domain [27], [28] or all given related domains [28]–[30]. The latter is often called multi-task metric learning (MTML), and is the focus of this paper. An implicit assumption of these methods is that the data samples of different domains lie in the same

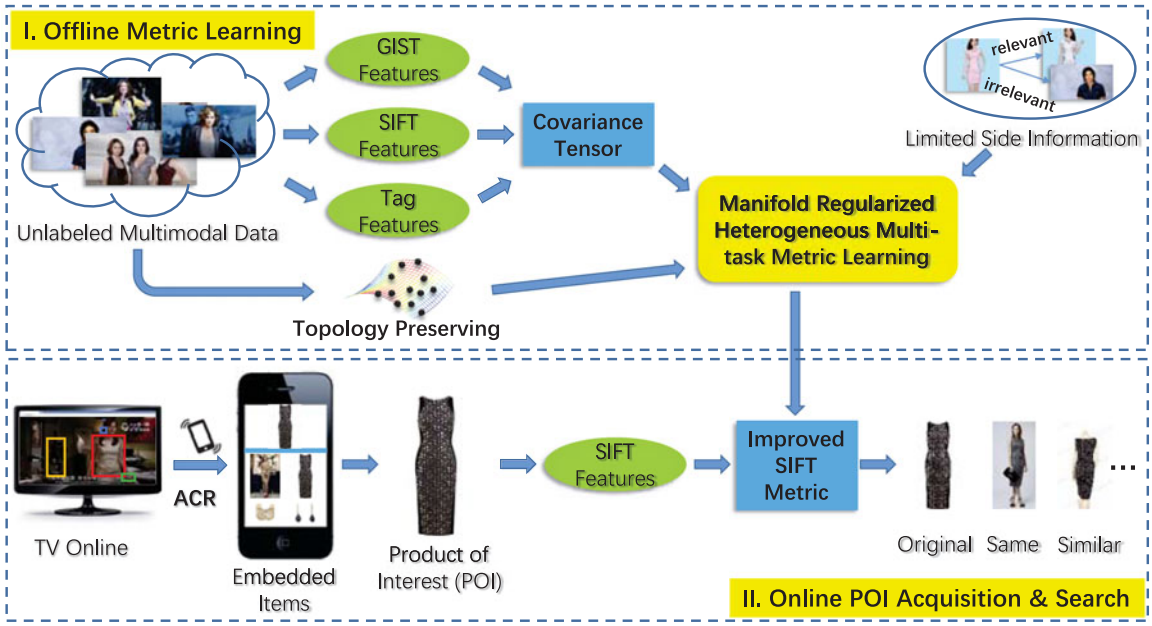


Fig. 1. Effortless TV-to-Online (T2O) architecture using the proposed manifold regularized heterogeneous multitask metric learning.

feature space, and so these approaches cannot handle heterogeneous features. Besides, these approaches utilize the "must-link" and "cannot-link" constraints, and thus may not work well for search. To remedy these drawbacks, we propose a novel manifold regularized heterogeneous MTML (MRHMTML) inspired by manifold regularization, ranking based DML and heterogeneous transfer learning.

B. Heterogeneous Transfer Learning

Developments in transfer learning across heterogeneous feature spaces can be grouped in two categories: heterogeneous domain adaptation (HDA) [17], [31] and heterogeneous multi-task learning (HMTL) [18]. In HDA, there is usually a single target domain that has limited labeled data, and our aim is to utilize the sufficient labeled data from related source domains to help the learning in the target domain. Whereas in HMTL, the labeled data in all domains are scarce, and thus we treat different domains equally and make them help each other.

Most HDA methods only incur two domains, i.e., one source and one target domain. The main idea in these methods is to either map the heterogeneous data into a common feature space by learning a feature mapping for each domain [32], [33], or map the data from the source domain to the target domain by learning an asymmetric transformation [31], [34]. The former is equivalent to Mahalanobis metric learning since each learned mapping could be used to derive a metric directly. Compared with HDA, there are much fewer works on HMTL, and one representative approach is the multi-task discriminant analysis (MTDA) [18], which extends linear discriminant analysis (LDA) to learn multiple tasks simultaneously by assuming a common intermediate structure is shared by the learned latent representations of different domains. MTDA can deal with more than two domains, but is limited in that only the pairwise correlations (between each latent representation and the shared representation) are exploited. Therefore, the high-order correlations between all domains are ignored in MTDA. This shortcoming is rectified in the proposed MRHMTML framework.

It is noted that heterogeneous multi-task DML method is different from multi-view DML [35], which is also used to deal with heterogeneous data. The goal of heterogenous multi-task DML method is to improve the performance of each DML task by utilizing the information of all different tasks, where the utilized features are different. However, multi-view (or multi-modal) DML is to learn an integrated distance metric by using all different features. In heterogeneous multi-task DML, the final prediction is performed in each domain based on the improved distance metric, where only a single type of feature is available. Distinctly in multi-view DML, features of all different domains should be provided in the prediction.

III. System Overview and Problem Formulation

Architecture of our developed TV-to-Online (T2O) system is presented first, and then the core module of product search, where a novel metric learning approach called MRHMTML is presented.

A. System Architecture

The T2O system is composed of two main parts as shown in Fig. 1: offline metric learning and online product of interest (POI) acquisition and search. In the first metric learning part, a large number of unlabeled products are collected, where both textual (such as tags, surrounding texts, etc.) and visual (i.e., image) information are contained. Different types of visual features are complementary with each other, such as local

bag of SIFT [12], and the global GIST [11]. They are extracted for image representation. Meanwhile, the textual information is also preprocessed for textual features extraction (such as TF-IDF [36]). We treat each feature space as a domain. To share the information across all different domains, we calculate a covariance tensor for all various features. In addition, we assume that each of these domains is provided with limited side information (in the form of relevance judgement to the query). By simultaneously using the limited side information, exploiting the high-order correlation information, and preserving the topology in each domain, we learn improved distance metric for each domain. The second part illustrates the process of online product search. When the users are watching TV, they can simply shake the mobile phone to capture the desired product item which is embedded in the TV shows. In this paper, we adopt the automatic content recognition (ACR) technique [37] for item acquisition. The product of interest (POI) can be specified easily by clicking it. In the T2O system, the specified POI is regarded as the input. We extract visual features, such as SIFT for the query. The most related items (such as the same, similar, or original styles) are retrieved for the query product from the database after the learned SIFT metric is utilized. At last, the system returns the retrieved products for users to shop. In our system, we assume the query has only one feature representation because: 1) usually, only visual information is contained in the query product from TV shows; 2) it is time-consuming to extract multiple features, and matching process is slow when the feature dimension is high. We give the technical details of the proposed MRHMTML below. Before that, we first summarize the used notations and concepts of multilinear algebra.

B. Notations

If $\mathcal{A}$ is an M-th order tensor of size $I_1 \times I_2 \times \ldots \times I_M$, and U is a $J_m \times I_m$ matrix, then the m-mode product of $\mathcal{A}$ and U is signified as $\mathcal{B} = \mathcal{A} \times_m U$, which is also an M-th order tensor of size $I_1 \times \ldots \times I_{m-1} \times J_m \times I_{m+1} \ldots \times I_M$ with the entry

$$\mathcal{B}(i_1, \ldots, i_{m-1}, j_m, i_{m+1}, \ldots, i_M) = \sum_{i_m=1}^{I_m} \mathcal{A}(i_1, i_2, \ldots, i_M) U(j_m, i_m). \quad (1)$$

The product of $\mathcal{A}$ and a set of matrices $\{U_m \in \mathbb{R}^{J_m \times I_m}\}_{m=1}^{M}$ is given by

$$\mathcal{B} = \mathcal{A} \times_1 U_1 \times_2 U_2 \ldots \times_M U_M. \quad (2)$$

The mode-m matricization of $\mathcal{A}$ is a matrix $A_{(m)}$ of size $I_m \times (I_1 \ldots I_{m-1} I_{m+1} \ldots I_M)$. We can regard the m-mode multiplication $\mathcal{B} = \mathcal{A} \times_m U$ as matrix multiplication in the form of $B_{(m)} = U A_{(m)}$.

Let $\mathbf{u}$ be an I_m-vector, the contracted m-mode product of $\mathcal{A}$ and $\mathbf{u}$ is denoted as $\mathcal{B} = \mathcal{A} \bar{\times}_m \mathbf{u}$, which is an $M-1$-th tensor of size $I_1 \times \ldots \times I_{m-1} \times I_{m+1} \ldots \times I_M$. The elements are calculated by

$$\mathcal{B}(i_1, \ldots, i_{m-1}, i_{m+1}, \ldots, i_M) = \sum_{i_m=1}^{I_m} \mathcal{A}(i_1, i_2, \ldots, i_M) \mathbf{u}(i_m). \quad (3)$$

Finally, the Frobenius norm of the tensor $\mathcal{A}$ is given by

$$\|\mathcal{A}\|_F^2 = \langle \mathcal{A}, \mathcal{A} \rangle = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \ldots \sum_{i_M=1}^{I_M} \mathcal{A}(i_1, i_2, \ldots, i_M)^2. \quad (4)$$

C. Problem Formulation

Suppose there are M heterogeneous domains, and the product database for the m'th domain is $\mathcal{D}_m = \{\mathbf{x}_{mi} \in \mathbb{R}^{d_m}, i = 1, \ldots, N_m\}$, and the corresponding query set is $Q_m = \{\mathbf{q}_{mi}, i = 1, \ldots, N_m^Q\}$. For each query $\mathbf{q}_{mi}$, we use $\mathcal{D}_{mi}^L = \{\mathbf{x}_{mi_j}, y_{mi_j}\}_{j=1}^{N_{mi}}$ to denote the set of labeled images with known relevance to $\mathbf{q}_{mi}$, and $y_{mi_j} \in \{+1, -1\}$ indicates $\mathbf{x}_{mi_j}$ is relevant to the query or not. Alternatively, we can use some initial distance metric to retrieve images for q_{mi} from $\mathcal{D}_m$, and choose the top returned images as $\mathcal{D}_{mi}^L$ [6]. We also assume that there are large amounts of unlabeled multi-domain products, i.e., $D^U = \{(\mathbf{x}_{1n}^U, \mathbf{x}_{2n}^U, \ldots, \mathbf{x}_{Mn}^U)\}_{n=1}^{N^U}$. That is, each product has feature representations in all domains. Then we have the following manifold regularized heterogeneous multi-task metric learning (MRHMTML) formulation,

$$\begin{aligned} \underset{\{A_m\}_{m=1}^M}{\arg\min} \; & F(\{A_m\}) = \sum_{m=1}^{M} \Psi(A_m) + \gamma R(A_1, A_2, \ldots, A_M) \\ & + \sum_{m=1}^{M} \gamma_m R_A(A_m), \\ \text{s.t. } & A_m \succeq 0, m = 1, 2, \ldots, M, \end{aligned} \quad (5)$$

where

$$\Psi(A_m) = \frac{1}{N_m'} \sum_{i=1}^{N_m^Q} \sum_{j,k=1}^{N_{mi}} L(A_m; \mathbf{q}_{mi}, \mathbf{x}_{mi_j}, \mathbf{x}_{mi_k}, y_{mi_j}, y_{mi_k})$$

and the loss for each training triplet $(\mathbf{q}_{mi}, \mathbf{x}_{mi_j}, \mathbf{x}_{mi_k})$ that satisfy $y_{mi_j} = +1, y_{mi_k} = -1$ is

$$\begin{aligned} & L(A_m; \mathbf{q}_{mi}, \mathbf{x}_{mi_j}, \mathbf{x}_{mi_k}, y_{mi_j}, y_{mi_k}) \\ & \quad = g([d(A_m; \mathbf{q}_{mi}, \mathbf{x}_{mi_k}) - d(A_m; \mathbf{q}_{mi}, \mathbf{x}_{mi_j})]). \end{aligned} \quad (6)$$

Here, $g(z) = \max\{0, 1 - z\}$ is the hinge loss, and N_m' is the number of triplets; $d(A_m; \mathbf{q}_{mi}, \mathbf{x}_{mi_j})$ is the distance between a query and database under the distance metric A_m. The definition is

$$d(A_m; \mathbf{q}_{mi}, \mathbf{x}_{mi_j}) = (\mathbf{q}_{mi} - \mathbf{x}_{mi_j})^T A_m (\mathbf{q}_{mi} - \mathbf{x}_{mi_j}). \quad (7)$$

We denote $\delta_{mi_j} = \mathbf{q}_{mi} - \mathbf{x}_{mi_j}$ for notational simplicity. Hence the distance can also be written as $d(A_m; \delta_{mi_j}) = \delta_{mi_j}^T A_m \delta_{mi_j}$. In addition, we signify $\delta_{ml}^+ = \delta_{mi_j}$ and $\delta_{ml}^- = \delta_{mi_k}$, where l is the triplet index and the corresponding query is $\mathbf{q}_{ml}$. The regularization term R is adopted to enforce knowledge being shared between different domains, and the regularization term

R_A is used to incorporate some prior knowledge or exploit some properties about the metric, such as sparse, low-rank, etc.

In this paper, we choose R_A to be a manifold regularization term [16], [38]. It is a popular regularization term used in semi-supervised learning, where there are only a few labeled samples but large amounts of unlabeled samples. To improve generalization ability of the model given the limited labeled data, geometry of the data distribution is exploited using the large amounts of unlabeled data and incorporated as a regularization term to penalize the model. The distribution is assumed to be supported on a low-dimensional manifold, which is approximated by the Laplacian of data adjacency graph. In this paper, we define the regularization as $R_A(A_m) = \sum_{i,j=1}^{N^U} w_{mij} d(A_m; \mathbf{x}_{mi}^U, \mathbf{x}_{mj}^U)$, where $w_{mij} = \exp(-\|\mathbf{x}_{mi}^U - \mathbf{x}_{mj}^U\|^2/(2\omega_m^2))$ is the edge weight between two neighboring nodes i and j in the data adjacency graph. Here, ω_m is a bandwidth hyper-parameter and empirically set as the mean of the Euclidean distances between all sample pairs, i.e., $\omega_m = \frac{1}{(N^U)^2}\sum_{i=1}^{N^U}\sum_{j=1}^{N^U}\|\mathbf{x}_{mi}^U - \mathbf{x}_{mj}^U\|^2$. If two samples are close to each in the original feature space, the weight will be large and thus their distance after applying the distance metric tends to be small when we minimize $R_A(A_m)$. In this way, topology of the data in each domain is preserved [17].

To enable knowledge transfer across domains, we decompose the matrix A_m as $A_m = U_m U_m^T$ according to the positive semi-definite property of the metric. Then we use the obtained transformation $U_m \in \mathbb{R}^{d_m \times r}$ to project the different representations of the multi-domain unlabeled data into a common subspace, where the correlation of all domains are maximized. This is formulated as the following problem:

$$\underset{\{U_m\}_{m=1}^M}{\arg\max} \frac{1}{N^U}\sum_{n=1}^{N^U} \mathrm{corr}(\mathbf{z}_{1n}^U, \mathbf{z}_{2n}^U, \ldots, \mathbf{z}_{Mn}^U), \tag{8}$$

where $\{\mathbf{z}_{mn}^U = U_m^T \mathbf{x}_{mn}^U\}_{m=1}^M$ are the projected representations of different domains for the n'th sample, and $\mathrm{corr}(\mathbf{z}_{1n}^U, \mathbf{z}_{2n}^U, \ldots, \mathbf{z}_{Mn}^U) = (\mathbf{z}_{1n}^U \odot \mathbf{z}_{2n}^U \odot \ldots \odot \mathbf{z}_{Mn}^U)^T \mathbf{e}$ is the correlation among all of them. Here, $\odot$ signifies the element-wise product, and $\mathbf{e} \in \mathbb{R}^r$ is a vector with all one elements. According to [39], the correlation can be rewritten as $\mathcal{G}\bar{\times}_1(\mathbf{x}_{1n}^U)^T \ldots \bar{\times}_M(\mathbf{x}_{Mn}^U)^T$, where $\mathcal{G} = \sum_{q=1}^r(\mathbf{u}_1^q \circ \mathbf{u}_2^q \circ \ldots \circ \mathbf{u}_M^q) = \mathcal{I}_r \times_1 U_1 \times_2 U_2 \ldots \times_M U_M$ is the covariance tensor of all transformations. Here, $\circ$ is the outer product, $\mathcal{I}_r \in \mathbb{R}^{r\times r\times\ldots\times r}$ is an identity tensor (the diagonal elements are 1, and all other entries are 0), and r is the number of common factors shared by all domains. Then the problem (8) is reformulated as

$$\underset{\{U_m\}_{m=1}^M}{\arg\max} \frac{1}{N^U}\sum_{n=1}^{N^U} \mathcal{G}\bar{\times}_1(\mathbf{x}_{1n}^U)^T \ldots \bar{\times}_M(\mathbf{x}_{Mn}^U)^T. \tag{9}$$

According to [40], we can reformulate the above problem as

$$\underset{\{U_m\}_{m=1}^M}{\arg\min} \frac{1}{N^U}\sum_{n=1}^{N^U} \|\mathcal{C}_n^U - \mathcal{G}\|_F^2, \tag{10}$$

where $\mathcal{C}_n^U = \mathbf{x}_{1n}^U \circ \mathbf{x}_{2n}^U \circ \ldots \circ \mathbf{x}_{Mn}^U$ is the covariance tensor of all the original feature representations for the n'th sample. The objective of (10) is rewrote as $\|\mathcal{C}^U - \mathcal{G}\|_F^2$ to accelerate computation. Here, $\mathcal{C}^U = \frac{1}{N^U}\sum_{n=1}^{N^U}\mathcal{C}_n^U$ is a sum of covariance tensor of all unlabeled samples. Then specific optimization problem for the proposed MRHMTML can be obtained by regarding $\|\mathcal{C}^U - \mathcal{G}\|_F^2$ as the regularizer $R(\cdot)$ in (5), i.e.,

$$\begin{aligned}\underset{\{U_m\}_{m=1}^M}{\arg\min}\ & F(\{U_m\}) \\ &= \sum_{m=1}^M \frac{1}{N_m'}\sum_{l=1}^{N_m'} g\left([d(U_m;\delta_{ml}^-) - d(U_m;\delta_{ml}^+)]\right) \\ &\quad + \gamma\|\mathcal{C}^U - \mathcal{G}\|_F^2 \\ &\quad + \sum_{m=1}^M \frac{\gamma_m}{(N^U)^2}\sum_{i,j=1}^{N^U} w_{mij}\|U_m^T\mathbf{x}_{mi}^U - U_m^T\mathbf{x}_{mj}^U\|_2^2, \\ &\text{s.t. } U_m \succeq 0, m = 1,2,\ldots,M,\end{aligned} \tag{11}$$

where $d(U_m;\delta_{ml}) = \delta_{ml}^T U_m U_m^T \delta_{ml}$, and the tradeoff hyper-parameters γ and $\{\gamma_m\}$ are positive. Non-negative relationship between the original feature representations are preserved by the non-negativity constraints $\{U_m \succeq 0\}$. It is intuitive that a latent subspace shared by all domains can be found by minimizing the second term in (11). In this subspace, the representations of different domains are close to each other and knowledge is transferred. Hence different domains can help each other to learn improved transformation U_m, and also the distance metric A_m.

IV. Optimization Algorithm

Problem (11) can be solved using an alternating optimization strategy. That is, only one variable U_m is updated at a time and all the other $U_{m'}$, $m' \neq m$ are fixed. This updating procedure is conducted iteratively for each variable. Following [41], we have

$$\mathcal{G} = \mathcal{I}_r \times_1 U_1 \times_2 U_2 \ldots \times_M U_M = \mathcal{B} \times_m U_m.$$

where $\mathcal{B} = \mathcal{I}_r \times_1 U_1 \ldots \times_{m-1} U_{m-1} \times_{m+1} U_{m+1} \ldots \times_M U_M$. According to the metricizing property of tensor, we have $G_{(m)} = U_m B_{(m)}$ and can easily verify that $\|\mathcal{C}^U - \mathcal{G}\|_F^2 = \|C_{(m)}^U - G_{(m)}\|_F^2$. This leads to the following sub-problem of (11) w.r.t. U_m:

$$\begin{aligned}\underset{U_m}{\arg\min}\ & F(U_m) = \Phi(U_m) + \Omega(U_m), \\ &\text{s.t. } U_m \succeq 0,\end{aligned} \tag{12}$$

where $\Phi(U_m) = \frac{1}{N_m'}\sum_{l=1}^{N_m'} g\left([d(U_m;\delta_{ml}^-) - d(U_m;\delta_{ml}^+)]\right)$, and $\Omega(U_m) = \gamma\|C_{(m)}^U - U_m B_{(m)}\|_F^2 + \frac{\gamma_m}{(N^U)^2}\mathrm{tr}(U_m^T X_m^U L_m (X_m^U)^T U_m)$. Here, $L_m = D_m - W_m$ is the graph Laplacian with the definition in [16], W_m is a weight matrix with the entry $W_m(i,j) = w_{mij}$, and D_m is a diagonal matrix with the element $D_m(i,i) = \sum_{j=1}^{N^U} W_m(i,j)$. The solution of problem (12) is found by employing the projected gradient method (PGM) developed in [42]. To apply PGM, we first smooth the non-differentiable terms in $\Phi(U_m)$ according to [43] for gradient calculation. In the following derivation, we omit the subscript m due to the reason of notational clarity. According

to the strategies suggested in [43], we smooth the hinge loss $g(U;q_l,\delta_l^-,\delta_l^+) = \max\{0,1-[d(U;\delta_l^-)-d(U;\delta_l^+)]\}$ as follows

$$g^\sigma(U;\mathbf{q}_l,\delta_l^-,\delta_l^+) = \max_{\nu\in\mathcal{Q}} \nu_l(1-[d(U;\delta_l^-)-d(U;\delta_l^+)]) - \frac{\sigma}{2}\|\mathbf{q}_l\|_\infty \nu_l^2, \tag{13}$$

where $\mathcal{Q}=\{\nu: 0\le\nu_l\le 1, \nu\in\mathbb{R}^{N'}\}$, and σ is the smooth hyper-parameter, where we set it as 0.5 empirically. It can be easily verified that the solution of ν_l in (13) is given by

$$\nu_l = \text{median}\left\{\frac{1-[d(U;\delta_l^-)-d(U;\delta_l^+)]}{\sigma\|\mathbf{q}_l\|_\infty}, 0, 1\right\}. \tag{14}$$

We obtain the following piece-wise approximation of g by substituting the above solution back into (13),

$$g^\sigma = \begin{cases} 0, & [d(U;\delta_l^-)-d(U;\delta_l^+)] > 1; \\ 1-[d(U;\delta_l^-)-d(U;\delta_l^+)] - \frac{\sigma}{2}\|\mathbf{q}_l\|_\infty, & [d(U;\delta_l^-)-d(U;\delta_l^+)] < 1-\sigma\|\mathbf{q}_l\|_\infty; \\ \frac{\left(1-[d(U;\delta_l^-)-d(U;\delta_l^+)]\right)^2}{2\sigma\|\mathbf{q}_l\|_\infty}, & \text{otherwise.} \end{cases} \tag{15}$$

Then the descent direction can be determined by computing gradient of the above smoothed hinge loss. The result is summarized in the following theorem.

Theorem 1: The gradient of the approximated hinge loss (15) is

$$\frac{\partial g^\sigma(U)}{\partial U} = \sum_{l=1}^{N'} \left(2\nu_l(\delta_l^+(\delta_l^+)^T - \delta_l^-(\delta_l^-)^T)U\right). \tag{16}$$

Here, ν_l is given in the form of (15), which is related to U.

The proof is presented in the appendix. Besides, it is easy to calculate that the gradient of $\Omega(U)$. Therefore, the gradient of the smoothed $F(U_m)$ is

$$\begin{aligned}\frac{\partial F^\sigma(U_m)}{\partial U_m} = &\frac{1}{N'_m}\sum_{l=1}^{N'_m}\left(2\nu_{ml}\left(\delta_{ml}^+(\delta_{ml}^+)^T - \delta_{ml}^-(\delta_{ml}^-)^T\right)U_m\right) \\ &+ 2\gamma\left(U_m B_{(m)}B_{(m)}^T - C_{(m)}^U B_{(m)}^T\right) \\ &+ \frac{2\gamma_m}{(N^U)^2}(X_m^U L_m (X_m^U)^T U_m),\end{aligned} \tag{17}$$

where ν_{ml} is given by (14).

After obtaining the gradient, we apply the improved PGM [42] to optimize the smoothed $F^\sigma(U_m)$, and the updating rule is

$$U_m^{t+1} = P[U_m^t - \mu_t \nabla F^\sigma(U_m^t)], \tag{18}$$

Here, $P[x]$ is an operator that projects the negative elements of x to zero, and the condition for choosing the step size μ_t is:

$$F^\sigma(U_m^{t+1}) - F^\sigma(U_m^t) \le \kappa \nabla F^\sigma(U_m^t)^T(U_m^{t+1}-U_m^t), \tag{19}$$

Following [42], we choose the hyper-parameter κ as 0.01. We determine the step size according to Algorithm 4 in [42], which has a guaranteed convergence property, and we refer to [42] for more details. The stopping condition is $|F^\sigma(U_m^{t+1}) - F^\sigma(U_m^t)|/(|F^\sigma(U_m^{t+1}) - F^\sigma(U_m^0)| < \epsilon)$, where we initialize U_m^0 as the results of the previous iterations in the alternating of all $\{U_m\}_{m=1}^M$.

By alternatively updating each U_m until convergence, i.e., $|OBJ_{k+1}-OBJ_k|/|OBJ_k| < \epsilon$, we obtain the solutions of (11). Here, OBJ_k is the objective value of (11) at the k'th iteration step. Our MRHMTML algorithm converge since in the alternating procedure, the objective value of (12) decreases at each step i.e., $F(U_m^{k+1}, \{U_{m'}^k\}_{m'\neq m}) \le F(\{U_m^k\})$. This indicates that $F(\{U_m^{k+1}\}) \le F(\{U_m^k\})$. After obtaining the solutions $\{U_m^*\}_{m=1}^M$, we derive the distance metric as $A_m^* = U_m^* U_m^{*T}$, which is utilized to improve distance estimation in the subsequent learning, such as product search in each domain (which is referred to feature space in this paper).

V. Complexity Analysis

To analyze the time complexity of the proposed MRHMTML algorithm, we first present the computational cost of optimizing each U_m, where the solution is found using the iterative PGM algorithm. In each iteration, we shall first determine the descent direction according to the gradient calculated using (17). Then an appropriate step size is obtained by exhaustedly checking whether the condition (19) is satisfied, where in each check we need to calculate the updated objective value of $F^\sigma(U_m^{t+1})$. To accelerate computation, we can pre-calculate $B_{(m)}B_{(m)}^T$, $C_{(m)}^U B_{(m)}^T$ and $X_m^U L_m (X_m^U)^T$, where the time costs are $O(r^2\prod_{m'\neq m} d_{m'})$, $O(r\prod_{m=1}^M d_m)$ and $O\left(\max(d_m(N^U)^2, d_m^2 N^U)\right)$ respectively. After the pre-calculation, the time complexity of calculating $(U_m^T U_m)(B_{(m)}B_{(m)}^T)$, $U_m^T(C_{(m)}^U B_{(m)}^T)$ and $U_m^T X_m^U L_m (X_m^U)^T U_m$ becomes $O(r^2 d_m + r^3)$, $O(r^2 d_m)$ and $O(r d_m^2)$ respectively. It is easy to derive that the computational cost of the remaining parts in the objective function is $O(r d_m N'_m)$. Considering that $r < d_m$, the time costs of calculating the objective value becomes $O(r d_m N'_m + r d_m^2)$. Similarly, we can derive that the time cost of calculating the gradient is also $O(r d_m N'_m + r d_m^2)$.

Therefore, the computational cost of optimizing U_m is $O[r\prod_{m'\neq m} d_{m'}(r+d_m) + \max(d_m(N^U)^2, d_m^2 N^U) + T_2 T_1 (r d_m N'_m + r d_m^2)]$, where T_1 is the number of required checks to find the step size, and T_2 is the number of iterations for reaching the stop criterion. Considering that the optimal rank $r \ll d_m$, we can simplify the cost as $O[r\prod_{m=1}^M d_m + \max(d_m(N^U)^2, d_m^2 N^U) + T_2 T_1 (r d_m N'_m + r d_m^2)]$. Finally, suppose the number of iterations for alternately updating all $\{U_m\}_{m=1}^M$ is Γ, we obtain the time complexity of the proposed MRHMTML, i.e., $O(\Gamma M[r\prod_{m=1}^M d_m + \max(\bar{d}_m(N^U)^2, \bar{d}_m^2 N^U) + T_2 T_1 (r\bar{d}_m \bar{N}'_m + r\bar{d}_m^2)])$, where $\bar{N}'_m$ and $\bar{d}_m$ are average number of labeled sample triplets and feature dimension of all domains respectively. This is linear w.r.t. M, $\bar{N}'_m$ and $\prod_{m=1}^M d_m$, and quadratic in the numbers r,

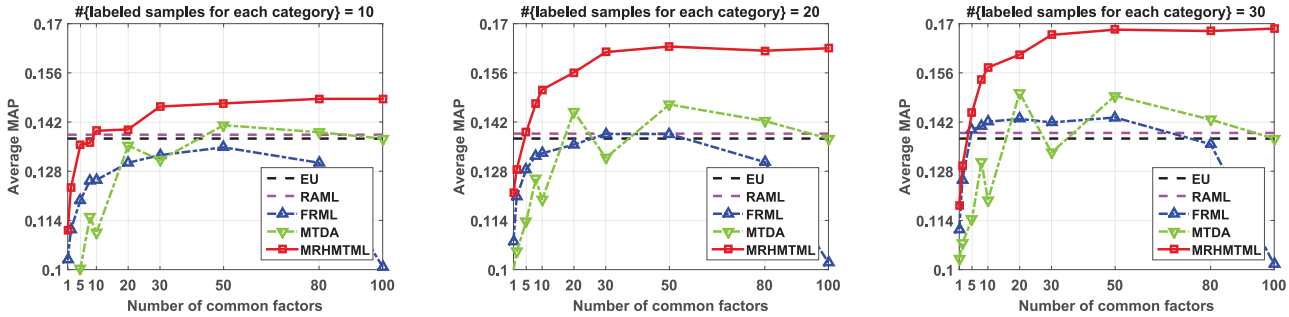


Fig. 2. Average MAP of all domains versus number of the common factors on the VOC dataset.

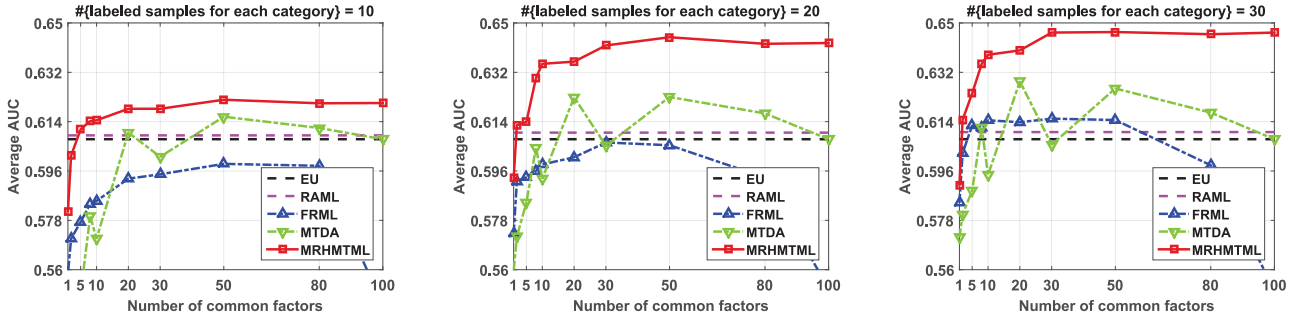


Fig. 3. Average AUC of all domains versus number of the common factors on the VOC dataset.

$\bar{d}_m$ and N^U. Besides, it is common that $\Gamma < 10$, $T_2 < 20$, and $T_1 < 50$, so the complexity is moderate.

VI. Experiments

In this section, we evaluate performance of the proposed MRHMTML in the object (product) search application. In the following, we first present the datasets to be used and experimental setups.

A. Datasets, Features, and Evaluation Criteria

We conduct the experiments on two public and challenging web image datasets, i.e., PASCAL VOC' 07 (VOC for short) [19] and NUS WIDE (NUS for short) [20].

The VOC dataset consists of around 10,000 images and 20 categories. We choose a subset of 13 product categories (e.g., chair, tv-monitor, dining-table, etc.) for experiments. There are 5,038 images in the resulting subset, and we use the features extracted by [44], in which a variety of visual features and tags are public available. From these features, we choose the popular SIFT [12] based local features, global GIST [11], and the tags as the different heterogenous domains. The feature dimensions are 1000, 512, and 804 respectively. We perform kernel PCA to preprocess these features to resolve comparable patterns for meaningful transfer. This can also reduce the running time. The resulting feature dimensions are all 100. The image set of each domain is split equally to form the training and test sets. We vary the number of labeled samples by randomly selecting $\{10, 20, 30\}$ for each category from the training set. The selected labeled ones are utilized to construct side information (in a triplet form $\langle query, relevance, irrelevance\rangle$) for distance metric learning. Each triplet is obtained by first selecting a labeled instance as query, and then its relevant and irrelevant sample are selected from the remaining labeled set according to whether the sample belongs to the same category of the query or not. The unlabeled set is an intersection of the remaining training data of all domains. For each concept, 20 queries are chosen for test.

In the NUS dataset, there are 269,648 images from 81 concepts, in which 10 of them can be regarded as products, e.g., computer, book, flower, etc. This results in a subset containing 23,539 images. The utilized features are the bag of local SIFT (500-D), global wavelet texture (WT, 128-D), and tag (1000-D). These features are provided by [20]. We vary the number of labeled instances for each concept in the set $\{6, 8, 10\}$, and all other settings are the same as VOC.

For both datasets, we perform search with distance estimation improved by the learned distance metric, and this is the task in each domain. The hyper-parameters are determined by conducting leave-one-out cross validation on the labeled training set. Two popular criteria, i.e., mean average precision (MAP) [45] and area under the ROC curve (AUC) [46] are adopted for evaluation. The following experiments are run five times by randomly choosing different sets of labeled samples.

B. Compared Methods

The methods included for comparison are:

- *EU:* directly using the simple Euclidean metric and original feature representations to compute the distance between samples.
- *RAML [6]:* a competitive ranking-based DML algorithm. The metric is learned separately for each domain. For this method, only the limited labeled samples are utilized in each domain, and no additional information (from other

TABLE I
AVERAGE MAP AND AUC OF ALL DOMAINS OF THE COMPARED METHODS AT THEIR BEST NUMBERS (OF COMMON FACTORS) ON THE VOC DATASET

	Average MAP			Average AUC		
Methods	10	20	30	10	20	30
EU	0.137 ± 0.000			0.608 ± 0.000		
RAML [6]	0.138 ± 0.002	0.139 ± 0.001	0.139 ± 0.001	0.609 ± 0.003	0.610 ± 0.002	0.610 ± 0.001
FRML [13]	0.135 ± 0.003	0.139 ± 0.005	0.143 ± 0.003	0.599 ± 0.003	0.606 ± 0.007	0.615 ± 0.003
MTDA [18]	0.141 ± 0.001	0.147 ± 0.001	0.150 ± 0.003	0.616 ± 0.002	0.623 ± 0.002	0.629 ± 0.003
MRHMTML	$\mathbf{0.149 \pm 0.001}$	$\mathbf{0.164 \pm 0.003}$	$\mathbf{0.169 \pm 0.002}$	$\mathbf{0.622 \pm 0.006}$	$\mathbf{0.645 \pm 0.004}$	$\mathbf{0.647 \pm 0.002}$

In each domain, the number of labeled training samples for each category varies from 10 to 30.

domains) is leveraged. We choose the trade-off hyper-parameter from the set $\{10^i | i = -5, -4, \ldots, 4\}$.

- *FRML [13]:* a recently proposed DML algorithm for ranking. The algorithm is quite efficient and scalable. We determine the hyper-parameter using the strategy in [13].
- *MTDA [18]:* a multi-task extension of the supervised dimension reduction technique LDA (linear discriminant analysis). The transforms $\{U_m\}$ are learned simultaneously for multiple heterogenous domains (feature spaces). The distance metric is derived as $A_m = U_m U_m^T$ after learning the transformation. We set the only hyper-parameter (intermediate dimensionality) as 100 empirically due to the insensitivity of the model to the hyper-parameter.
- *MRHMTML:* the proposed manifold regularized heterogeneous multi-task metric learning algorithm for ranking. We set the hyper-parameters γ_m as the same value, and tune both γ and γ_m over the set $\{10^i | i = -5, -4, \ldots, 4\}$.

A common subspace is learned in both MTDA and the proposed MRHMTML. Determination of r, which is the number of common factors (dimensionality of the common subspace) is still an open problem. We thus report the performance on a variety of $r \in \{1, 2, 5, 8, 10, 20, 30, 50, 80, 100\}$, which is also applied to the metric rank in FRML.

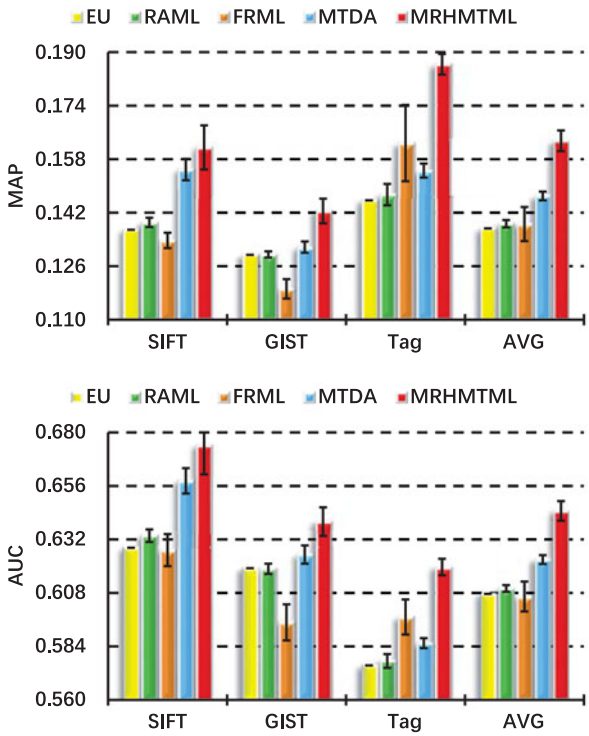


Fig. 4. Individual MAP and AUC score of each domain of the compared methods at their best numbers (of common factors) on the VOC dataset (20 labeled instances for each category; AVG: average).

C. Evaluation on the VOC Dataset

1) Average Performance: We show the average performance (MAP and AUC score) of all domains in Figs. 2 and 3. In Table I, we summarize the peak performance of different methods, where both the mean and standard variation are reported. It can be observed from these results that: 1) when more labeled instances are given, all of the compared methods tend to achieve better performance; EU is kept unchanging since it is pre-defined and does not make use of the label information in search; 2) when comparing with the EU baseline, the improvements of single-task DML algorithms (RAML and FRML) are only slight. FRML is even worse than EU when the number of labeled samples for each category is 10. The main reason is that they learn the metrics for different domains separately, and thus it is hard for them to achieve satisfactory performance given the limited number of labeled samples; 3) in contrast, performance of the heterogeneous multi-task approaches (MTDA and MRHMTML) are much better than EU. Therefore, leveraging information from other domains can be very useful in DML; Usually, the optimal r is less than 30. Hence we may only need 30 factors to distinguish the different categories in this dataset; 4) the proposed MRHMTML outperforms MTDA in most cases and the performance curve is more smooth. This can be interpreted as the expressive ability of the factors learned by our method are stronger than MTDA. This may be because that the high-order correlations of all domains are exploited in our method, while MTDA only discover the pairwise correlations between domains; 5) the performance of our method in terms of AUC is in consistent with MAP, and there is a significant 11.6% relative improvements over MTDA under the MAP criterion (20 labeled samples for each category).

2) An Investigation on Individual Domain: Performance of each domain at the best number of common factors are shown in Fig. 4. We can see from the results that: 1) RAML and the EU baseline are comparable, and FRML is only superior to EU

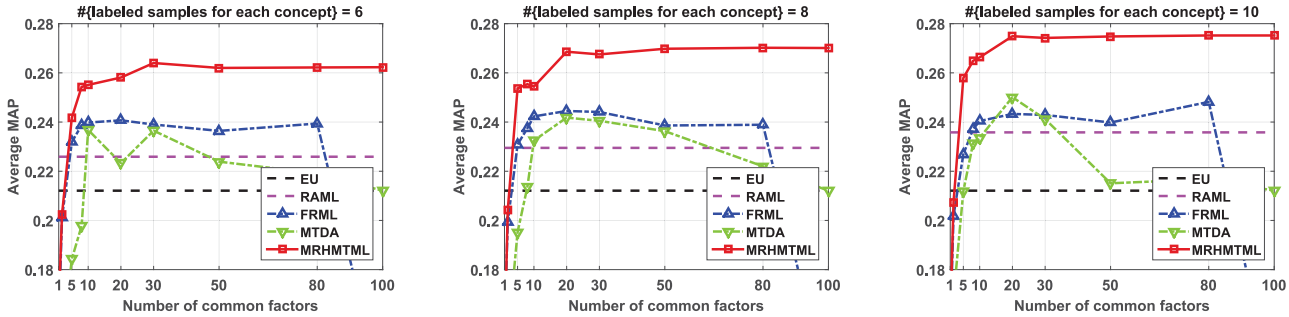


Fig. 5. Average MAP of all domains versus number of the common factors on the NUS dataset.

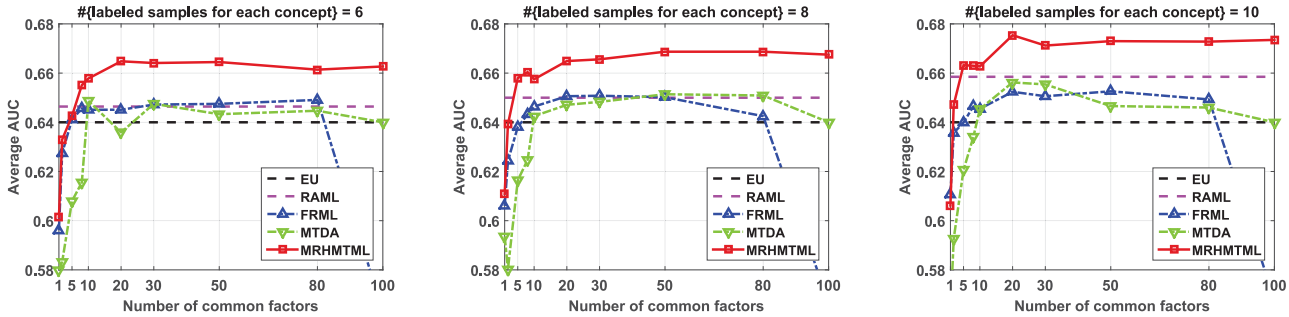


Fig. 6. Average AUC of all domains versus number of the common factors on the NUS dataset.

TABLE II
AVERAGE MAP AND AUC OF ALL DOMAINS OF THE COMPARED METHODS AT THEIR BEST NUMBERS (OF COMMON FACTORS) ON THE NUS DATASET

	Average MAP			Average AUC		
Methods	6	8	10	6	8	10
EU	0.212 ± 0.000			0.640 ± 0.000		
RAML [6]	0.226 ± 0.010	0.230 ± 0.012	0.236 ± 0.009	0.646 ± 0.005	0.650 ± 0.006	0.659 ± 0.006
FRML [13]	0.241 ± 0.003	0.245 ± 0.005	0.248 ± 0.003	0.649 ± 0.006	0.651 ± 0.007	0.653 ± 0.002
MTDA [18]	0.237 ± 0.009	0.242 ± 0.004	0.250 ± 0.005	0.649 ± 0.005	0.651 ± 0.007	0.656 ± 0.005
MRHMTML	$\mathbf{0.264 \pm 0.004}$	$\mathbf{0.270 \pm 0.005}$	$\mathbf{0.275 \pm 0.002}$	$\mathbf{0.665 \pm 0.005}$	$\mathbf{0.669 \pm 0.004}$	$\mathbf{0.675 \pm 0.005}$

In each domain, the number of labeled training samples for each concept varies from 6 to 10.

in the tag domain. The main reason is that the side information provided for training in each domain are scarce and the different domains do not communicate with each other. In contrast, each domain is improved by the multi-task methods. This indicates that different domains successfully help each other in learning the metrics by transforming knowledge across them; 2) MTDA is better than the proposed MRHMTML in only one domain in terms of MAP, while MRHMTML outperforms MTDA consistently in terms of AUC. This demonstrates that our method can well discover the high-order correlation information, and this is better than exploiting paired correlation information in MTDA.

D. Evaluation on the NUS Dataset

1) Average Performance: The MAP and AUC scores of different methods are shown in Figs. 5 and 6 respectively. Table II is a summarization of the peak performance (results at the best numbers of common factors). From the results, we observe that: 1) the single-task DML algorithms (RAML and FRML) take effect on this dataset, and significant improvements are obtained when comparing with the baseline (EU). The main reason may be that separability of the different concepts are larger that of categories in the VOC dataset. Thus the side information are more discriminative; 2) the multi-task MTDA is only comparable to and sometime even worse than the single-task approaches. This is mainly because both RAML and FRML adopt the ranking-based loss, while MTDA is not designed for ranking. Moreover, only the pairwise correlation information is exploited in MTDA. In contrast, we obtain satisfactory results since ranking-based loss is adopted and high-order relationships of all domains are explored. The tendency of the AUC and MAP score curves are consistent. Thus the superiority of our method is further verified.

2) Sensitivity Analysis w.r.t. the Hyperparameters: We show the performance w.r.t. different choices of the hyper-parameters γ and γ_m in Fig. 7. From the results, we can see that: 1) the best performance is achieved when both of the hyper-parameters are neither too large nor too small. Therefore, both of the introduced regularization terms $R(A_1, A_2, \ldots, A_M)$ and

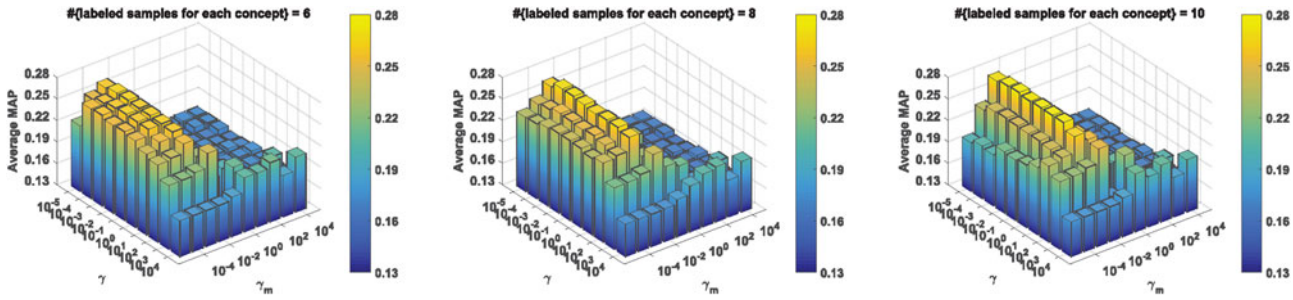


Fig. 7. Performance w.r.t. different choices of the hyperparameters on the NUS dataset.

$R_A(A_m)$ play critical roles in obtaining optimal metrics for different domains; 2) the optimal hyper-parameters are similar for different number of labeled samples. This demonstrates stability of the proposed model.

VII. Conclusion

An effective TV-to-Online (T2O) system aims to make it easier for people to shop online while watching TV. This paper introduces a novel transfer distance metric learning algorithm to address the distance estimation problem, which plays a vital role in products matching module of a T2O system. The proposed method takes full advantage of multiple domains (feature spaces) by analyzing their feature covariance tensor. In addition, we exploit the geometric structure of the data to make full use of the unlabeled data and employ ranking-based loss to make the learned metric especially appropriate and feasible to match similar products.

The main conclusions of the experiments on two challenging and popular datasets are: 1) a separate metric learning for each domain may degrade performance if the side information is given insufficiently. Meanwhile, the deficiency problem of labeled data can be alleviated if the metrics of multiple heterogeneous domains are learned simultaneously. This result is consistent with description in the literatures for multi-task learning [47]; 2) transfer learning methods can exploit the shared knowledge across different domains. The high-order statistics (correlation information) play a critical role in discovering appropriate common factors, which can benefit each domain; 3) the ranking-based loss is adopted to help learn an efficient metric for products matching. Despite these advantages, a flaw in the proposed algorithm is that the limited side information must be provided for all domains. In the future, we intend to design some algorithm for such case that only one domain is provided with the side information. There exist some approaches that can annotate products in videos [48] or learn concept (e.g., product) relationships [49] for visual search. Incorporate these techniques into our system may further improve our product search performance. Moreover, we aim to make a web-based data collection from online shopping sites to create a large scale product dataset. Relying on the huge amount of potential users, we anticipate that T2O would be an emerging trend that will greatly facilitate customer shopping. Our approach would make great contribution to transform this process in both TV and online video market.

Appendix A
Proof of Theorem 1

Proof: According to (14) and (15), we can calculate the gradient of g^σ for the l'th example as

$$\frac{\partial g^\sigma(U;\mathbf{q}_l,\delta_l^-,\delta_l^+)}{\partial U} = \begin{cases} 0, & \nu_l = 0; \\ 2(\delta_l^+(\delta_l^+)^T - \delta_l^-(\delta_l^-)^T)U, & \nu_l = 1; \\ 2\nu_l(\delta_l^+(\delta_l^+)^T - \delta_l^-(\delta_l^-)^T)U, & \nu_l = \frac{1-[d(U;\delta_l^-)-d(U;\delta_l^+)]}{\sigma\|\mathbf{q}_l\|_\infty}. \end{cases} \tag{20}$$

This indicates that

$$\frac{\partial g^\sigma(U;\mathbf{q}_l,\delta_l^-,\delta_l^+)}{\partial U} = 2\nu_l(\delta_l^+(\delta_l^+)^T - \delta_l^-(\delta_l^-)^T)U. \tag{21}$$

Thus the sum of the gradient over all the N' examples is

$$\frac{\partial g^\sigma(U)}{\partial U} = \frac{\partial \sum_{l=1}^{N'} g^\sigma(U;\mathbf{q}_l,\delta_l^-,\delta_l^+)}{\partial U} = \sum_{l=1}^{N'} \left(2\nu_l(\delta_l^+(\delta_l^+)^T - \delta_l^-(\delta_l^-)^T)U\right). \tag{22}$$

Here, ν_l is given by (14) and thus related to U. This completes the proof. ■

Acknowledgment

The authors would like to thank the handling associate editor and all the anonymous reviewers for their constructive comments.

References

[1] Y. Wen, X. Zhu, J. J. Rodrigues, and C. W. Chen, "Cloud mobile media: Reflections and outlook," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 885–902, Jun. 2014.

[2] "Micro-moments: Your guide to winning the shift to mobile," 2015. [Online]. Available: https://www.thinkwithgoogle.com/micromoments/intro.html

[3] Y. Jin, Y. Wen, and H. Hu, "Minimizing monetary cost via cloud clone migration in multi-screen cloud social TV system," in *Proc. IEEE Global Commun. Conf.*, 2013, pp. 1747–1752.

[4] I. Gonzalez-Diaz, M. Birinci, F. Diaz-de Maria, and E. J. Delp, "Neighborhood matching for image retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 544–558, Mar. 2017.

[5] S. Karaoglu, R. Tao, T. Gevers, and A. W. Smeulders, "Words matter: Scene text for image classification and retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1063–1076, May 2017.

[6] J.-E. Lee, R. Jin, and A. K. Jain, "Rank-based distance metric learning: An application to image retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2008, pp. 1–8.

[7] E. P. Xing, M. I. Jordan, S. Russell, and A. Ng, "Distance metric learning with application to clustering with side-information," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2002, pp. 505–512.

[8] H. Wang, L. Feng, J. Zhang, and Y. Liu, "Semantic discriminative metric learning for image similarity measurement," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1579–1589, Aug. 2016.

[9] Y. Luo, T. Liu, D. Tao, and C. Xu, "Decomposition-based transfer distance metric learning for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3789–3801, Sep. 2014.

[10] Z.-C. Song and S.-G. Liu, "Sufficient image appearance transfer combining color and texture," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 702–711, Apr. 2017.

[11] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[12] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] D. Lim and G. Lanckriet, "Efficient learning of Mahalanobis metrics for ranking," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1980–1988.

[14] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[15] Y. Luo, Y. Wen, and D. Tao, "Heterogeneous multitask metric learning across multiple domains," *IEEE Trans. Neural Netw. Learn. Syst.*, Oct. 4, 2017, doi: 10.1109/TNNLS.2017.2750321.

[16] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, no. 11, pp. 2399–2434, 2006.

[17] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1541–1546.

[18] Y. Zhang and D.-Y. Yeung, "Multi-task learning in heterogeneous feature spaces," in *Proc. AAAI Conf. Artif. Intell.*, 2011, pp. 574–579.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge 2007 (VOC2007) results," 2007.

[20] T.-S. Chua *et al.*, "NUS-WIDE: A real-world web image database from national university of singapore," in *Proc. Int. Conf. Image Video Retrieval*, 2009.

[21] B. Kulis, "Metric learning: A survey," *Found. Trends Mach. Learn.*, vol. 5, no. 4, pp. 287–364, 2012.

[22] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *arXiv preprint arXiv:1306.6709v4*, 2014.

[23] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. 17th Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 513–520.

[24] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2005, pp. 1473–1480.

[25] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 209–216.

[26] B. McFee and G. R. Lanckriet, "Metric learning to rank," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 775–782.

[27] B. Geng, D. Tao, and C. Xu, "DAML: Domain adaptation metric learning," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2980–2989, Oct. 2011.

[28] Y. Zhang and D.-Y. Yeung, "Transfer metric learning with semi-supervised extension," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, 2012, Art. no. 54.

[29] S. Parameswaran and K. Q. Weinberger, "Large margin multi-task metric learning," in *Proc. 23rd Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1867–1875.

[30] P. Yang, K. Huang, and C.-L. Liu, "Geometry preserving multi-task metric learning," *Mach. Learn.*, vol. 92, no. 1, pp. 133–175, 2013.

[31] J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan, "Heterogeneous domain adaptation for multiple classes," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2014, pp. 1095–1103.

[32] X. Shi, Q. Liu, W. Fan, P. S. Yu, and R. Zhu, "Transfer learning on heterogenous feature spaces via spectral transformation," in *Proc. Int. Conf. Data Mining*, 2010, pp. 1049–1054.

[33] L. Duan, D. Xu, and I. Tsang, "Learning with augmented features for heterogeneous domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 711–718.

[34] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 1785–1792.

[35] P. Xie and E. P. Xing, "Multi-modal distance metric learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1806–1812.

[36] K. Aas and L. Eikvil, "Text categorization: A survey," Norwegian Comput. Center, Oslo, Norway, Tech. Rep., 1999.

[37] R. Suzuki, D. Sakamoto, and T. Igarashi, "AnnoTone: Record-time audio watermarking for context-aware video editing," in *Proc. ACM Conf. Human Factors Comput. Syst.*, 2015, pp. 57–66.

[38] Y. Luo *et al.*, "Multiview vector-valued manifold regularization for multilabel image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 709–722, May 2013.

[39] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen, "Tensor canonical correlation analysis for multi-view dimension reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3111–3124, Nov. 2015.

[40] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[41] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors," *SIAM J. Matrix Analysis Appl.*, vol. 21, no. 4, pp. 1324–1342, 2000.

[42] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, 2007.

[43] Y. Nesterov, "Smooth minimization of non-smooth functions," *Math. Program.*, vol. 103, no. 1, pp. 127–152, 2005.

[44] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2010, pp. 902–909.

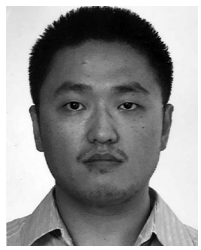
[45] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A support vector method for optimizing average precision," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2007, pp. 271–278.

[46] T. Joachims, "A support vector method for multivariate performance measures," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 377–384.

[47] Y. Luo, Y. Wen, D. Tao, J. Gui, and C. Xu, "Large margin multi-modal multi-task feature extraction for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 414–427, Jan. 2016.

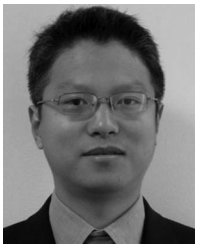
[48] G. Li, M. Wang, Z. Lu, R. Hong, and T.-S. Chua, "In-video product annotation with web information mining," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 8, no. 4, 2012, Art. no. 55.

[49] R. Hong, Y. Yang, M. Wang, and X.-S. Hua, "Learning visual semantic relationships for efficient visual retrieval," *IEEE Trans. Big Data*, vol. 1, no. 4, pp. 152–161, Dec. 2015.



Qiang Fu received the master's degree from Northwestern Polytechnical University, Xi'an, China, in 2007. He is currently working toward the Ph.D. degree in engineering with Peking University, Beijing, China. He has more than ten years' industrial experience in streaming media services and solutions. He and his team used to innovate the first multiscreen P2P media system in China in 2007 and China's first smart TV middleware framework for DVB+OTT service in 2011. His research interests include multimedia systems, machine learning, cloud computing, and embedded devices.

Yong Luo received the B.E. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2009, and the D.Sc. degree in electronics engineering and computer science from Peking University, Beijing, China, in 2014. He is currently a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He was a Visiting Student with the School of Computer Engineering, Nanyang Technological University, and the Faculty of Engineering and Information Technology, University of Technology Sydney. He has authored several scientific articles at top venues including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the International Join Conference on Artificial Intelligence (IJCAI), and the AAAI Conference on Artificial Intelligence. His research interests primarily include machine learning and data mining with applications to visual information understanding and analysis. Mr. Luo was a recipient of the IEEE Globecom 2016 Best Paper Award and was nominated as the IJCAI 2017 Distinguished Best Paper Award.



Yonggang Wen (S'99–M'08–SM'14) received the Ph.D. degree in electrical engineering and computer science (with a minor in western literature) from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008. He is currently an Associate Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He has been with Cisco, San Jose, CA, USA, where he led product development in content delivery network, which had a revenue impact of $3 billion globally. His work in multiscreen cloud social TV has been featured by global media (over 1600 news articles from over 29 countries). He has authored or coauthored more than 140 papers in top journals and prestigious conferences. His current research interests include cloud computing, green data centers, big data analytics, multimedia networks, and mobile computing. Prof. Wen was a recipient of the ASEAN ICT Award 2013 (Gold Medal) and the Data Center Dynamics Awards 2015-APAC for his work on cloud 3-D view, as the only academia entry, and a co-recipient of the 2015 IEEE Multimedia Best Paper Award and the Best Paper Award at the EAI/ICST Chinacom 2015, the 2014 International Conference on Wireless Communications and Signal Processing, the IEEE Globecom 2013, and the 2012 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing. He was the Chair for the IEEE ComSoc Multimedia Communication Technical Committee in 2014–2016.

Dacheng Tao (F'15) is a Professor of computer science and ARC Laureate Fellow with the School of Information Technologies and the Faculty of Engineering and Information Technologies, and the Inaugural Director of the UBTECH Sydney Artificial Intelligence Centre, University of Sydney, Darlington, NSW, Australia. He mainly applies statistics and mathematics to Artificial Intelligence and Data Science. His research results have expounded in one monograph and more than 500 publications in prestigious journals and prominent conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the *Journal of Machine Learning Research*, the *International Journal of Computer Vision*, the Conference on Neural Information Processing Systems, the International Conference on Machine Learning, the Conference on Computer Vision and Pattern Recognition, the International Conference on Computer Vision, the European Conference on Computer Vision, the IEEE International Conference on Data Mining (ICDM), and the ACM SIGKDD Conference. His research interests include computer vision, data science, image processing, machine learning, and video surveillance. Prof. Tao has been a recipient of several best paper awards, such as the Best Theory/Algorithm Paper Runner Up Award at IEEE ICDM 2007, the Best Student Paper Award at IEEE ICDM 2013, the Distinguished Student Paper Award at the 2017 International Joint Conferences on Artificial Intelligence, the 2014 ICDM 10-Year Highest-Impact Paper Award, and the 2017 IEEE Signal Processing Society Best Paper Award. He was a recipient the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award, and the 2015 UTS Vice-Chancellor's Medal for Exceptional Research. He is a Fellow of the American Association for the Advancement of Science, the Optical Society of America, the International Association for Pattern Recognition, and the International Society for Optical Engineers.

Ying Li is currently a Professor with the School of Software and Microelectronics, Peking University, Beijing, China. From 2001 to 2012, she was a Senior Manager leading the Department of Distributed Computing, IBM China Research Center. She has built leading-edge technology and made solid contributions to several IBM commercial distributed software systems and solutions. She has authored or coauthored more than 50 academic papers in international journal and conferences and filed more than 30 patents in distributed computing area. She served as a PC member of several international conferences and reviewer of international journals. Her research interests include automatic computing and distributed systems.

Ling-Yu Duan (M'06) received the Ph.D. degree in information technology from the University of Newcastle, Callaghan, NSW, Australia, in 2008. He is currently a Full Professor with the National Engineering Laboratory of Video Technology, School of Electronics Engineering and Computer Science, Peking University (PKU), Beijing, China. He has served as the Associate Director of the Rapid-Rich Object Search Laboratory, a joint laboratory between Nanyang Technological University, Singapore, and PKU, since 2012. Before he joined PKU, he was a Research Scientist with the Institute for Infocomm Research, Singapore, from March 2003 to August 2008. His research interests include multimedia indexing, search, and retrieval, mobile visual search, visual feature coding, and video analytics. His recent major achievements have focused on the topic of compact representation of visual features and high-performance image search. Prof. Duan was a recipient of the *EURASIP Journal on Image and Video Processing* Best Paper Award in 2015, the Ministry of Education Technology Invention Award (First Prize) in 2016, the National Technology Invention Award (Second Prize) in 2017, the China Patent Award for Excellence (2017), and the National Information Technology Standardization Technical Committee "Standardization Work Outstanding Person" Award in 2015. He was a co-Editor of MPEG Compact Descriptor for Visual Search Standard (ISO/IEC 15938-13). He is a co-Chair of MPEG Compact Descriptor for Video Analytics.