UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

LEARNING WITH SHARED INFORMATION FOR IMAGE AND VIDEO ANALYSIS

Gaowen Liu

Advisor

Prof. Nicu Sebe

Università degli Studi di Trento

2017

Publications

This thesis consists of the following publications:

- Chapter 2:
  Gaowen Liu, Yan Yan, Ramanathan Subramanian, Jingkuan Song, Guoyu Lu, Nicu Sebe: Active Domain Adaptation with Noisy Labels for Multimedia Analysis. World Wide Web 19(2): 199-215, 2016

- Chapter 3:
  Gaowen Liu, Yan Yan, Elisa Ricci, Yi Yang, Yahong Han, Stefan Winkler, Nicu Sebe: Inferring Painting Style with Multi-Task Dictionary Learning. International Joint Conference on Artificial Intelligence (IJCAI): 2162-2168, 2015
  Gaowen Liu, Yan Yan, Jingkuan Song, Nicu Sebe: Minimizing dataset bias: Discriminative multi-task sparse coding through shared subspace learning for image classification. International Conference on Image Processing (ICIP): 2869-2873, 2014

- Chapter 4:
  Yan Yan, Elisa Ricci, **Gaowen Liu**, Nicu Sebe: Egocentric Daily Activity Recognition via Multitask Clustering. IEEE Transactions on Image Processing 24(10): 2984-2995, 2015

The following are the papers published during the course of the Ph.D but not included in this thesis:

- **Gaowen Liu**, Yan Yan, Chenqiang Gao, Wei Tong, Alexander G. Hauptmann, Nicu Sebe:The Mystery of Faces: Investigating Face Contribution for Multimedia Event Detection. ACM International Conference on Multimedia Retrieval (ICMR), 2014

- Yan Yan, Elisa Ricci, Ramanathan Subramanian, **Gaowen Liu**, Oswald Lanz, Nicu Sebe: A Multi-Task Learning Framework for Head Pose Estimation under Target Motion. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(6): 1070-1083, 2016

- Yan Yan, Yi Yang, Deyu Meng, **Gaowen Liu**, Wei Tong, Alexander G. Hauptmann, Nicu Sebe: Event Oriented Dictionary Learning for Complex Event Detection. IEEE Transactions on Image Processing 24(6): 1867-1878, 2015

- Yan Yan, Yi Yang, Haoquan Shen, Deyu Meng, **Gaowen Liu**, Alexander G. Hauptmann, Nicu Sebe: Complex Event Detection via Event Oriented Dictionary Learning. AAAI Conference on Artificial Intelligence (AAAI): 3841-3847, 2015

- Yan Yan, Haoquan Shen, **Gaowen Liu**, Zhigang Ma, Chenqiang Gao, Nicu Sebe: GLocal tells you more: Coupling GLocal structural for feature selection with sparsity for image and video classification. Computer Vision and Image Understanding, vol 124: 99-109, 2014

- Yan Yan, Elisa Ricci, Ramanathan Subramanian, **Gaowen Liu**, Nicu Sebe: Multitask Linear Discriminant Analysis for View Invariant Action Recognition. IEEE Transactions on Image Processing 23(12): 5599-5611, 2014

- Yan Yan, Elisa Ricci, **Gaowen Liu**, Ramanathan Subramanian, Nicu Sebe: Clustered Multi-task Linear Discriminant Analysis for View Invariant Color-Depth Action Recognition. International Conference on Pattern Recognition (ICPR): 3493-3498, 2014

- Chenqiang Gao, Deyu Meng, Wei Tong, Yi Yang, Yang Cai, Haoquan Shen, **Gaowen Liu**, Shicheng Xu, Alexander G. Hauptmann:Interactive Surveillance Event Detection through Mid-level Discriminative Represen-

tation. ACM International Conference on Multimedia Retrieval (ICMR), 2014

- Yan Yan, **Gaowen Liu**, Elisa Ricci, Nicu Sebe: Multi-task linear discriminant analysis for multi-view action recognition. International Conference on Image Processing (ICIP): 2842-2846, 2013

- Yan Yan, Zhongwen Xu, **Gaowen Liu**, Zhigang Ma, Nicu Sebe: GLocal structural feature selection with sparsity for multimedia data understanding. ACM Multimedia Conference (MM): 537-540, 2013

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Image and video recognition is a fundamental and challenging problem in computer vision, which has progressed tremendously fast recently. In the real world, a realistic setting for image or video recognition is that we have some classes containing lots of training data and many classes that contain only a small amount of training data. Therefore, how to use the frequent classes to help learning the rare classes is an open question. *Learning with shared information* is an emerging topic which can solve this problem. There are different components that can be shared during concept modelling and machine learning procedure, such as sharing generic object parts, sharing attributes, sharing transformations, sharing regularization parameters and sharing training examples, *etc*. For example, representations based on attributes define a finite vocabulary that is common to all categories, with each category using a subset of the attributes. Therefore, sharing some common attributes for multiple classes will benefit the final recognition system.

In this thesis, we investigate some challenging image and video recognition problems under the framework of *learning with shared information*. My Ph.D research comprised of two parts. The first part focuses on the two domains (source and target) problems where the emphasis is to boost the recognition performance on the target domain by utilizing useful knowledge from source

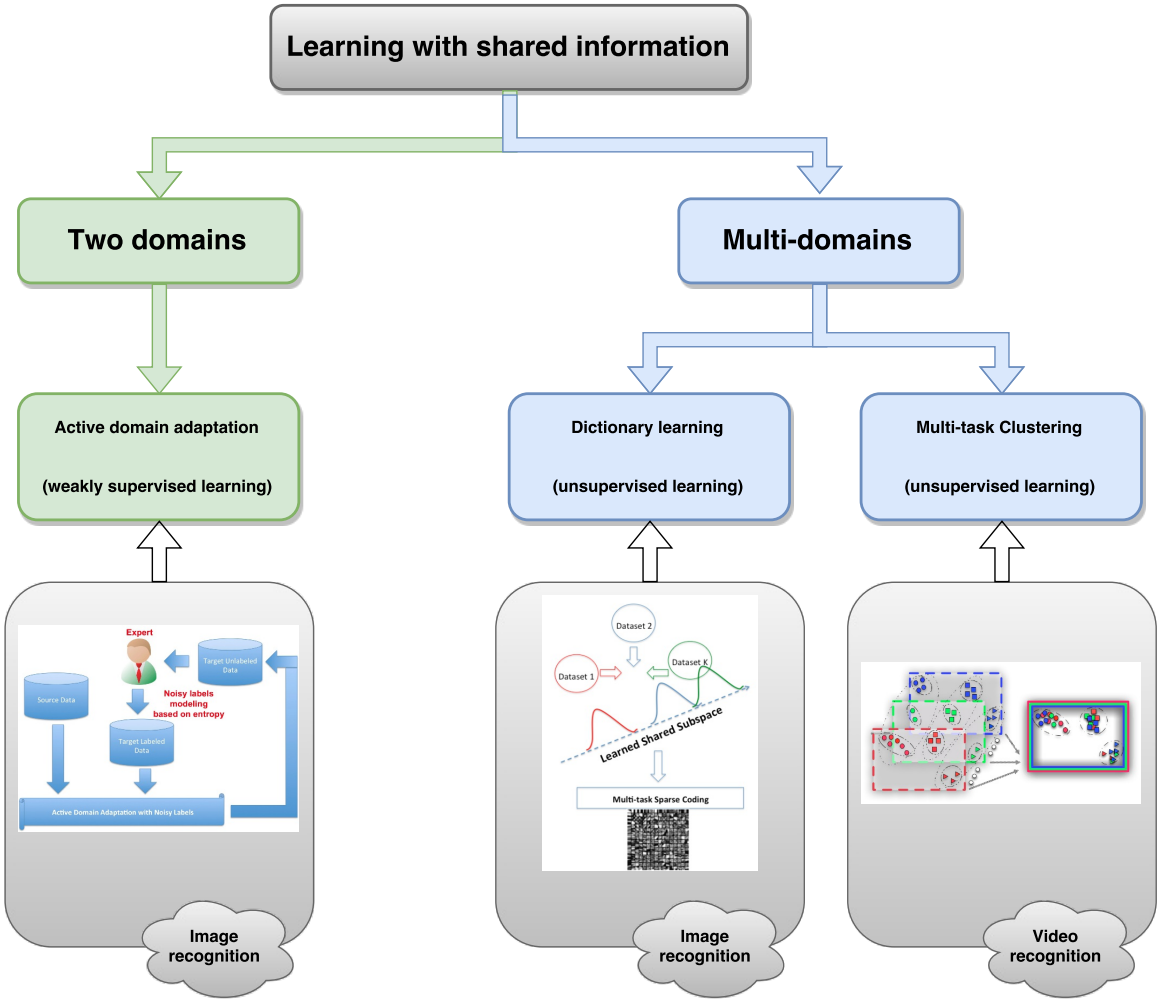


Figure 1.1: The overview of this thesis.

domain. The second part focuses on multi-domains problems where all domains are considered equally important. This means we want to improve performance for all domains by exploring the useful information across domains. Fig.1.1 shows the overview of this thesis. These two parts can be summarized as *learning with shared information*. In particular, we investigate three topics to achieve this goal in the thesis, which are active domain adaptation, multi-task learning, and dictionary learning, respectively.

1.1 Contribution of the Thesis

To sum up, this thesis makes the following contributions towards *learning with shared information*:

- For two domains (source and target) image recognition problem, an active domain adaptation framework is proposed to illustrate the power of learning with shared information.
- For multiple domains (parallel tasks), we introduce the idea of learning style-specific dictionaries. A novel multi-task dictionary learning is proposed for automatic analysis of painting image recognition.
- For multiple domains (parallel tasks), we also propose an unsupervised approach, a multi-task clustering framework, for the first-person vision activity video recognition.

1.2 Overview of the Thesis

In Chapter 2, we begin our research from image recognition problem using domain adaptation which involves transferring useful information from source domain to the target domain to boost recognition performance. Moreover, we integrate domain adaptation (DA) with active leaning (AL) which helps to minimize the effort for the acquisition of labelled data. We perform extensive experiments on different datasets to evaluate our strategy.

Multi-task learning is another approach for learning with shared information. Multi-task learning is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias. It does this by learning tasks in parallel while using a shared representation. What is learned for each task can help other tasks be learned better. In Chapter 3, we propose a new multi-task dictionary learning

approach by uncovering a shared subspace of different datasets. We also investigate a multi-task dictionary learning approach for inferring painting styles. The results show that multi-task dictionary learning achieve better performance for image recognition.

To step further, we move on our research on discovering and extracting relevant patterns from videos. In Chapter 4, we consider the videos collected from wearable cameras of several people performing daily activities. We notice that, videos continuously record several hours of human life. The data is heterogeneous and labelling them is an intensive and boring task requiring extensive human labour. In order to tackle the problem, an unsupervised multi-task clustering framework is proposed for video activity analysis in Chapter 4.

In summary, Chapters 2, 3, 4 present three different strategies to perform learning with shared information. Chapter 5 presents the conclusions and the future research directions.

Chapter 2

Active Domain Adaption[1]

Supervised learning methods require sufficient labelled examples to learn a good model for classification or regression. However, available labelled data are insufficient in many applications. Active learning (AL) and domain adaptation (DA) are two strategies to minimize the required amount of labelled data for model training. AL requires the domain expert to label a small number of highly informative examples to facilitate classification, while DA involves tuning the source domain knowledge for classification on the target domain. In this chapter, we demonstrate how AL can efficiently minimize the required amount of labelled data for DA. Since the source and target domains usually have different distributions, it is possible that the domain expert may not have sufficient knowledge to answer each query correctly. We exploit our active DA framework to handle incorrect labels provided by domain experts. Experiments with multimedia data demonstrate the efficiency of our proposed framework for active DA with noisy labels.

[1]Gaowen Liu, Yan Yan, Ramanathan Subramanian, Jingkuan Song, Guoyu Lu, Nicu Sebe: Active Domain Adaptation with Noisy Labels for Multimedia Analysis. World Wide Web 19(2): 199-215, 2016

2.1 Introduction

In machine learning, supervised methods require sufficient labelled examples in order to learn a good model. However, it is difficult to acquire sufficient labelled data in many real world applications. Moreover, labelling is an intensive task requiring extensive human labor. In order to tackle this problem, several approaches have been proposed. *Semi-supervised learning* aims to exploit the consistency between labelled and unlabelled data for classification. *Active learning* (AL) focuses on selecting a small set of essential examples for querying labels from domain experts. *Domain adaptation* (DA, also called *Transfer learning*) facilitates classification when the training (source) and test (target) data are from different domains. Domain adaptation uses the knowledge acquired from a large number of labelled source examples and a few labelled target examples for classification in the target domain.

DA algorithms (see [76] for a survey) seek to combine limited target data with the source data in order to adapt to the target domain. However, they typically tend to choose target examples *randomly* without considering which samples are most informative for classification in the target domain. Therefore, one question that needs to be examined is whether and how we can efficiently label target data for DA? Considering that the goal of both domain adaptation and active learning is to minimize labor-intensive data labelling, it would be worthwhile to integrate DA and AL in a single framework.

To our knowledge, very few works studied how to minimize the amount of labelled target data, especially under noisy labelling. A theoretical study on the number of labelled examples required to learn all targets to achieve an arbitrarily specified accuracy is presented in [118]. Two active transfer learning algorithms that allow for changes in all marginal and conditional distributions with the additional assumption that these changes are smooth are proposed in [100]. However, they do not consider noisy labels which are likely to occur in active

DA scenarios. [88] propose active transfer learning, but their approach is limited by the unlikely assumption that identical prediction labels are generated for a target example by the out-of-domain (source) and in-domain (target) classifiers. Additionally, the error rate of the transfer classifier is not bounded, and only binary classification is considered here. Extending active transfer learning to multi-class classification as in this work, the upper-bound error rate increases considerably and consequently, the domain-adaptive classifier cannot classify correctly anymore. In this chapter, we investigate an adaptive DA algorithm within an AL framework able to cope with label noise. We also extend the binary classification to a multi-class classification problem through error-correcting output coding. We investigate how AL helps to minimize the numbers of labelled data for DA even under noisy labelling. Experiments on real-world datasets for head-pose estimation and image classification demonstrate the efficacy of our proposed framework. To sum up, this chapter makes the following contributions:

- An active domain adaptation framework under noisy labelling is proposed, and is shown to be effective for multimedia analysis;

- We integrate active learning with domain adaptation for a multi-class setting through error correcting output coding;

- The proposed framework is general, and potentially applicable to many multimedia problems.

The chapter is organized as follows. Section 2.2 reviews related work from the perspective of active learning, domain adaptation and learning with noisy labels. Section 2.3 details active domain adaptation with noisy labels. Section 2.4 presents experimental results on head pose estimation and image classification, while Section 2.5 concludes the chapter.

2.2 Related Work

In this section, we review related work in the areas of active learning, domain adaptation and learning with noisy labels.

2.2.1 Active Learning

Active learning (AL) involves asking the domain expert to label a small number of most-informative examples to facilitate classification. Based on query scenarios, AL can be divided into three types of settings: (i) Membership query synthesis, (ii) stream-based selective sampling and (iii) pool-based sampling. The pool-based scenario has been studied for many real-world problems in machine learning and computer vision. Uncertainty sampling is a common approach in AL. Distance from hyperplane for margin-based classifiers has been used as a measure of uncertainty in previous works. [96] provided a theoretical motivation for SVM-based AL using the notion of a version space. [103] proposed a unified multi-class AL approach through error-correcting output coding based on the 'best worst case', which approximates the expected loss function with the smallest loss function among all the possible labels.

[44] extended the Fisher information framework to the batch-mode setting for binary logistic regression. [87] studied the problem of using several heuristics that take into account estimates of both oracle and model-uncertainty, and showed that data can be improved by selective repeated labelling. However, their analysis assumed both were equally and consistently noisy and annotation was a noisy process over some underlying true label. [59] introduced a novel criterion that requested a partial ordering for a set of examples that minimized the total rank margin in attribute space, subject to a visual diversity constraint.

Existing AL strategies can have uneven performance, being efficient on some datasets but ineffective on others, or inconsistent just between runs on the same dataset. [3] proposed perplexity-based graph construction and a new hierar-

chical sub-query evaluation algorithm to combat this variability and to use the potential of expected error reduction. [32] developed an efficient active learning framework based on convex programming, which can select multiple samples at a time for annotation. Unlike the state-of-the-art, their algorithm can be used in conjunction with any classifier type, including sparsity-based classifiers (SRC). [45] presented a collaborative computational model for AL with multiple human oracles. This approach leads not only to an ensemble kernel machine robust to noisy labels, but also to a principled label-quality measure detecting irresponsible labellers online.

[58] presented a novel multi-level AL approach to reduce the human annotation effort for training robust scene classification models. Different from most existing AL methods that can only query labels for selected instances at the class level, their approach established a semantic framework that predicted scene labels based on a latent object-based image representation, and was capable of querying labels at two different levels– the scene-class level and the latent object-class level. [120] proposed a semi-supervised batch mode multi-class AL algorithm for visual concept recognition. [18] proposed a novel convex, semi-supervised multi-label feature selection algorithm applicable to large-scale datasets.

2.2.2 Domain Adaptation

Traditional machine learning algorithms are based on the assumption that training and test data share the same distribution in feature space. When the training and test distributions are different, the classification accuracy drops significantly. In such cases, domain adaptation (DA) between the two domains is desirable. DA assumes that the training and testing data could be from different domains and distributions. It is motivated by the fact that people can intelligently apply knowledge learned previously to solve new problems efficiently. The target of DA is to find some common property which is shared between

the training (or source) and test (or target) domains. [76] identified three main research issues in DA: (i) what to transfer, (ii) how to transfer, and (iii) when to transfer. 'What to transfer' examines which knowledge can be transferred across domains or tasks. After discovering which knowledge can be transferred, learning algorithms are developed to describe the process of 'how to transfer'. 'When to transfer' studies the situations where the knowledge could be transferred, in order to guard against negative knowledge transfer that could hurt classification performance on the target domain.

There are several DA approaches. *Instance-transfer* involves re-weighting some source data for use in the target domain under the assumption that source data can be reused in the target domain ([24, 46, 124]). *Feature-representation-transfer* attempts to find a 'good' feature representation that reduces the difference between the source and target domains as well as the classification/regression error ([6, 26]). *Parameter-transfer* involves discovery of shared parameters or priors between the source and target models which can benefit from transfer learning ([10, 33, 79]). *Relational-knowledge-transfer* builds a mapping of relational knowledge between the source and target domains ([72]).

In essence, transfer learning adapts useful source information to efficiently classify in the target domain whose attributes vary with respect to the source. [26] proposed a feature replication method to augment features for transfer learning. [84] and [53] proposed a method for domain adaption using metric learning by generating cross-domain constraints. [24] used a boosting framework ([37]) to re-weight the importance of source and target samples for DA. [124] extended the transfer boosting framework to include information from multiple sources. [116] adapted DA by learning a delta function between the source and target domains based on SVMs. This method seeks the target decision boundary which is close to the source decision boundary. [29] extended this method via multiple kernel learning by learning kernels that minimize the mismatch between source and target domains. [43] proposed a framework for

image attribute adaptation. [128] proposed a DA framework for still-to-motion Adaptation (SMA) for human action recognition. [41] proposed finding a low-dimensional optimal consensus representation from multiple heterogeneous features for multi-view transfer learning. [42] proposed a sparse multi-label learning method to circumvent the visually polysemous barrier of multiple tags.

2.2.3 Learning with Noisy Labels

Nowadays, with the exponential growth of user-generated web images and videos, there has been an increasing interest in learning models that can handle noisy labels for supervised learning. This is a practical problem due to the uncontrolled environments in which humans label data. Given the importance of learning from noisy labels, a great deal of progress has been made in this regard. [73] addressed the problem of risk minimization in the presence of random noise, and obtained generalizable results using unbiased estimators and weighted loss functions. Efficient algorithms were proposed with both methods with provable guarantees for learning under label noise. [121] proposed a multimedia retrieval framework based on semi-supervised ranking and relevance feedback. [115] proposed event-oriented dictionary learning for multimedia event detection. [9] investigated the robustness of SVMs under adversarial label noise and proposed an improved method based on kernel matrix correction.

In active learning, it is highly probable that the expert may have no information concerning some queries and cannot provide accurate labels. [28] studied AL under noisy labelling with a human-like oracle by introducing non-uniformly distributed noise. They made a realistic assumption that the less confident the oracle is in labelling the example, the larger is the effect of the noise. [89] proposed a pool-based active learning framework through robust measures based on density power divergence. By minimizing β-divergence and γ-divergence, one can estimate the model accurately even with noisy labels. [38] tackled the fundamental problem of Bayesian active learning with noise,

where they needed to adaptively select from a number of expensive tests in order to identify an unknown hypothesis sampled from a known prior distribution. Learning with noisy labels is especially important in DA scenarios. To the best of our knowledge, there is no work focusing on active transfer learning with noisy labels.

2.3 Active Domain Adaptation with Noisy Labels

Domain adaptation uses a small number of labelled samples from the target domain. However, taking into account that not all samples from the target domain are equally informative, an efficient sample selection strategy is preferable. To minimize the amount of labelled data in the target domain, we attempt AL using different sample selection strategies.

2.3.1 SVM-based Domain Adaptation

Recently, several adaptation methods for the support vector machine classifier (SVM) were proposed for video retrieval in [29]. In order to make the SVM classifier adaptive to a new domain, the target decision function $f^T(x)$ is formulated as:

$$f^T(x) = f^S(x) + \Delta f(x) \tag{2.1}$$

where x is the specific feature vector and $f^S(x)$ is the source decision function. $\Delta f(x)$ is the function of the mismatch between source and target domains.

[29] extended this method via multiple kernel learning. In this case, the target decision function is formulated as:

$$f^T(x) = \sum_{p=1}^{P} \gamma_p f_p(x) + \sum_{m=1}^{M} d_m w_m^T \phi_m(x) + b \tag{2.2}$$

where $f_p(x)$ is the p-th pre-learned classifier trained using labelled data from both domains. P is the number of pre-learned classifiers. γ_p are the coeffi-

cients of the p-th pre-learned classifier. A linear combination of multiple kernels $\sum_{m=1}^{M} d_m w_m^T \phi_m(x) + b$ is used to model $\Delta f(x)$ in this setting with a bias term b. M is the number of kernels and d_m are the coefficients of the m-th kernel. w_m^T is the transpose of the weight vector w_m and $\phi_m(x)$ is the nonlinear feature mapping function where base kernels can be calculated as $k_m(x_i, x_j) = \phi_m^T(x_i)\phi_m(x_j)$.

There are two objectives to minimize. The first objective is to reduce the mismatch between the source and target domains. [39] proposed a similarity measure for two different distributions. The mismatch is measured by Maximum Mean Discrepancy (MMD) as in [46] based on the distance between the sample means from the source and target domains in the Reproducing Kernel Hilbert Space (RKHS) namely:

$$DIST(D^S, D^T) = \Omega(d) = \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \phi(x_i^S) - \frac{1}{n_T} \sum_{i=1}^{n_T} \phi(x_i^T) \right\|_H \tag{2.3}$$

where x_i^S and x_i^T are the samples from the source and target domains, respectively. n_S and n_T are the number of samples in the source and target domains. The second objective is to minimize the structural risk functional $J(d)$ in the target domain. If we combine these two objectives, the optimization problem is given by

$$\min_d G(d) = \frac{1}{2}\Omega^2(d) + \theta J(d) \tag{2.4}$$

where d is coefficient vector for the multiple kernels. $\Omega^2(d)$ is the distance between the source and target distributions. By introducing Lagrangian multipliers α, the dual form of the optimization is:

$$J(d) = \max_\alpha \alpha^T - \frac{1}{2}(\alpha y)^T \left(\sum_{m=1}^{M} d_m \widetilde{K_m}\right)(\alpha y) \tag{2.5}$$

This is equivalent to the dual form of SVM with kernel matrix $\sum_{m=1}^{M} d_m \widetilde{K_m}$, where

$\widetilde{K_m}$ are the domain-adaptive rectified kernels. The optimization problem can be solved by an existing SVM solver, such as LIBSVM ([17]).

2.3.2 Multiclass Active Learning

Margin-based learning algorithms minimize the loss function $L(\cdot)$ with respect to the margin.

$$\min \frac{1}{m} \sum_{i=1}^{m} L(y_i f(x_i)) \tag{2.6}$$

[2] proposed a unifying framework for studying the solution of multi-class categorization problems by reducing them to multiple binary problems. Firstly, we define a *coding matrix* $M \in \{-1, 0, +1\}^{k \times l}$. k is the number of classes and l is the number of binary classification problems. Let $M(r)$ denote the row r of M and $f(x)$ be the vector of predictions on an instance x, $f(x) = (f_1(x), ..., f_l(x))$. The basic idea is to predict with the label r, which row in $M(r)$ is the closest to the prediction $f(x)$, *i.e.*, predict label r that minimizes the distance $d(M(r), f(x))$. Taking advantage of the confidence of binary predictions, [2] proposed a loss-based decoding scheme. The idea is to choose the label r that is the most consistent with the predictions $f_s(x)$ in the sense that, if the example x was labelled r, the total loss on example (x, r) would be minimized over choices of $r \in Y$. The distance measure is the total loss on a proposed example (x, r).

$$d_L(M(r), f(x)) = \sum_{s=1}^{l} L(M(r, s) f_s(x)) \tag{2.7}$$

The predicted label $\hat{y} \in \{1, ..., k\}$ is:

$$\hat{y} = \arg\min_r d_L(M(r), f(x)) \tag{2.8}$$

[103] proposed an approximated sample selection strategy which uses the *best worst case* model to approximate the expected loss function with the small-

est loss function among all the possible labels.

$$\arg\max_{x} \min_{y \in Y} \sum_{s=1}^{l} L(M(y, s) f_s(x)) \tag{2.9}$$

If y_x is the predicted label for example x, Eqn.(2.9) becomes:

$$\arg\max_{x} \sum_{s=1}^{l} L(M(y_x, s) f_s(x)) \tag{2.10}$$

Here, we choose the most ambigous examples with the maximum expected loss for the predicted label.

2.3.3 Modeling with Noisy Labels

Information-theoretic methods can be used to model expert labelling knowledge. In the traditional AL scenario, the expert is able to provide a label for each queried instance. Then, the objective of uncertainty sampling based AL is to query the instance with the highest entropy. We model the domain expert as either knowledgeable to label an instance or not knowledgeable. The Knowledge Base (N) is defined as the union of instances (N^+) which have been labelled by the domain expert, and those instances (N^-) which the domain expert is unable to label.

The expected entropy of an unlabelled instance x_i with respect to sets N^+ and N^- is given by:

$$E = P(x_i \in N^+) E(y_i | x_i \in N^+) + P(x_i \in N^-) E(y_i | x_i \in N^-)$$

where $E(\cdot)$ is the entropy of samples x_i with respect to the predicted classifier label. Moreover, in the above equation $E(y_i | x_i \in N^-) = 0$ due to the definition of conditional entropy. The diverse density concept proposed in [69] is adopted to estimate $P(x_i) \in N^+$.

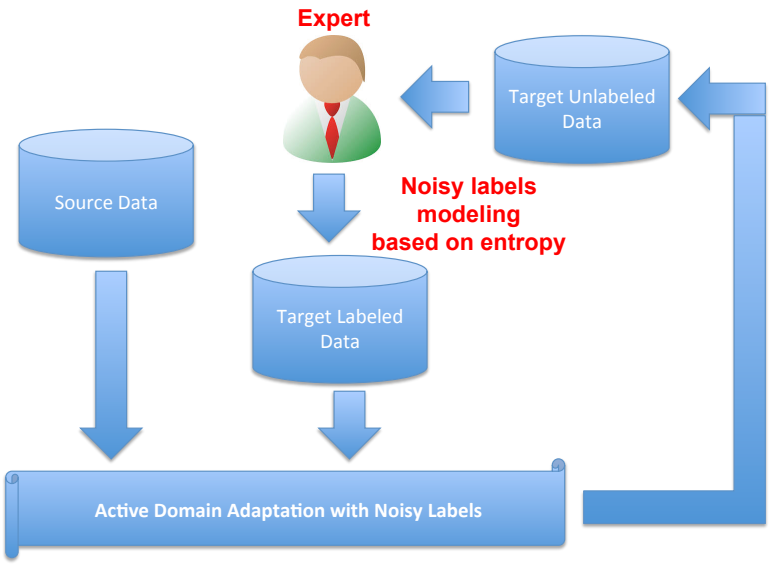


Figure 2.1: Framework for Active Domain Adaptation with Noisy Labels: labelled source, labelled and unlabelled target data are used to train the transfer classifier. Active learning is performed to select unlabelled target data to be labelled by the expert.

2.3.4 Framework

Considering that the goal of both DA and AL is to minimize intensive data labelling, it is reasonable to investigate how combining them can further minimize data labelling on the target. We propose an active DA under noisy labelling framework as shown in Fig.2.1. We use labelled source, labelled and unlabelled target data to train the transfer classifier. Then, we use AL to select unlabelled target data to be labelled by the expert, and add the same to labelled target data to update the transfer classifiers.

Algorithm 1 presents the active DA under noisy labelling algorithm. We initially randomly select one sample per category. Steps (4-8) represent the DA procedure. We combine labelled target samples D_l^s with labelled source samples D_l^t to train an adaptive SVM classifier $f^{T^m}(x)$ on the target domain D^t. To this end, we employ alternative coordinate descent to optimize variables α and d in Eqn.(2.5). η_t is the learning rate and g_t denotes the update direction. We iterate this procedure T_{max} times. Steps (9-12) represent the AL procedure.

In step (9), we calculate loss values for all the unlabelled target samples. We choose those unlabelled target samples that produce the least loss to be labelled by experts, and then add these samples to the labelled target domain. Steps (13-19) represent the procedure adopted to deal with noisy labels. If the expert does not know the label for x_i, the algorithm will include x_i in the negative knowledge base (N^-). Step (19) is to update the knowledge base N. We iterate this procedure K times.

Algorithm 1: Active Domain Adaptation under Noisy Labelling.

1 **Input:** Labelled *source* data D^s and unlabelled *target* data D^t. Let $D^t = D^t_l \cup D^t_u$. Randomly label one *target* sample per class and add them to D^t_l.
2 **Output:** Target sample label.

3 **for** $k = 1, ..., K$ do
4 Perform domain adaptation on D^t using samples from $D^s_l \cup D^t_l$ to obtain $f^{T^m}(x)$.
5 • **for** $t = 1, ..., T_{max}$ do
6 • Solve dual SVM variable α_t using LIBSVM with kernel matrix $\sum_{m=1}^{M} d_m \widetilde{K_m}$.
7 • Update the base kernel coefficients d_t by $d_{t+1} = d_t - \eta_t g_t$.
8 • **end for**
9 For all the samples $x_i \in D^t_u$, calculate loss function $\arg\max_x \sum_{s=1}^{l} L(M(y_x, s) f_s(x))$.
10 For all the samples $x_i \in D^t_u$, estimate $P(x_i) \in N^+$, then calculate expected entropy of x_i.
11 Select samples s^* according to the sum of least loss and expected entropy.
12 Get label y_{s^*}.
13 **if** the expert *knows* the label **then**
14 Add sample $s^* = (x_{s^*}, y_{s^*})$ to D^t_l.
15 $N^+ \leftarrow N^+ \cup x_i$.
16 **else**
17 $N^- \leftarrow N^- \cup x_i$.
18 **end if**
19 $N \leftarrow N^- \cup N^+$ and update knowledge N.
20 Classification using $f^{T^m}(x)$ on target domain test data.
21 **end for**

2.4 Experiments

In this section, we test the proposed active DA method for cross-domain headpose estimation (proposed earlier in [106]) and cross-domain web image classification (proposed in [84]).

2.4.1 Cross-domain Headpose Dataset

In video surveillance, knowing *where a person is looking at* is important. However, headpose estimation or classification from surveillance videos can be very hard, due to the low resolution and noise characterizing the sensor data. We focus on headpose estimation from low-resolution images acquired using a multi-camera system.

The CLEAR 2007 dataset ([93]) illustrated in Fig.2.2(a) provides multi-view images, output from four cameras placed in the room's corners. This dataset includes 15 persons rotating in-place, and exhibiting all possible head orientations while wearing a magnetic motion sensor (flock-of-birds) to measure their head pose. The task is to estimate the person's 3D head orientation with respect to the room's coordinate system, and to obtain a robust, joint pose estimate from all four views instead of employing only a single camera view for analysis.

In order to evaluate cross-domain headpose classification, we used the DPOSE dataset (described in [79]) shown in Fig.2.2(b). DPOSE is recorded under the same settings as CLEAR, with both static and moving persons (only data corresponding to static persons are used in our experiments). As evident from Fig.2.2, the illumination and recording environments are very different in the CLEAR and DPOSE datasets.

We firstly localize the head in each of the four views using the procedure described in [79]. The localized head regions are then resized to 20×20 resolution. We then concatenate the head crops from the four views on which visual features are extracted. Head pan is divided into eight classes, each denoting a 45°

(a) (b)

Figure 2.2: 4-view exemplar from the (a) CLEAR and (b) DPOSE datasets. Automatically extracted face crops are shown on the bottom right inset.

pan range, and for each head pan range, the tilt is quantized into three classes– namely *frontal* [-20°, 20°], *upward* (20°, 90°] and *downward* (-20°, -90°]. This leads to a total of 24 headpose classes (*e.g.* pan range (-22.5,22.5) with *frontal*, *upward* and *downward* tilts denote headpose classes 1–3). We divide the 4-view head image into 25 patches (every patch is 4×4). For the visual features computed over each view, we use HOG (81 dimensions) and skin pixel histograms (25 dimensions denoting the number of skin pixels in each patch). Then, we concatenate these features to derive a 106-dimensional vector per view, and a 424-dimension vector over the 4-view image.

We use several baseline methods to evaluate and compare our transfer learning results. $S_A T_B$ means we train on source domain A and test on target domain B. $S_B T_B$ means we train on target domain B and test on B. $S_{(A+B)} T_B$ means we train on both A and B and test on B. $TrAdaboost$ means we use the Adaboost algorithm ([37]) trained on labelled source and target data. AMKL_random means we use adaptive multiple kernel learning and randomly label target samples. AMKL_active (our method) means we use AMKL and actively label the target samples. For all the experiments, we report the mean accuracy on 5 randomly selected train/test sets. SVM parameter $C = 1$ in all the experiments. We use 100 images per class in the source domain and query 24 samples (one sam-

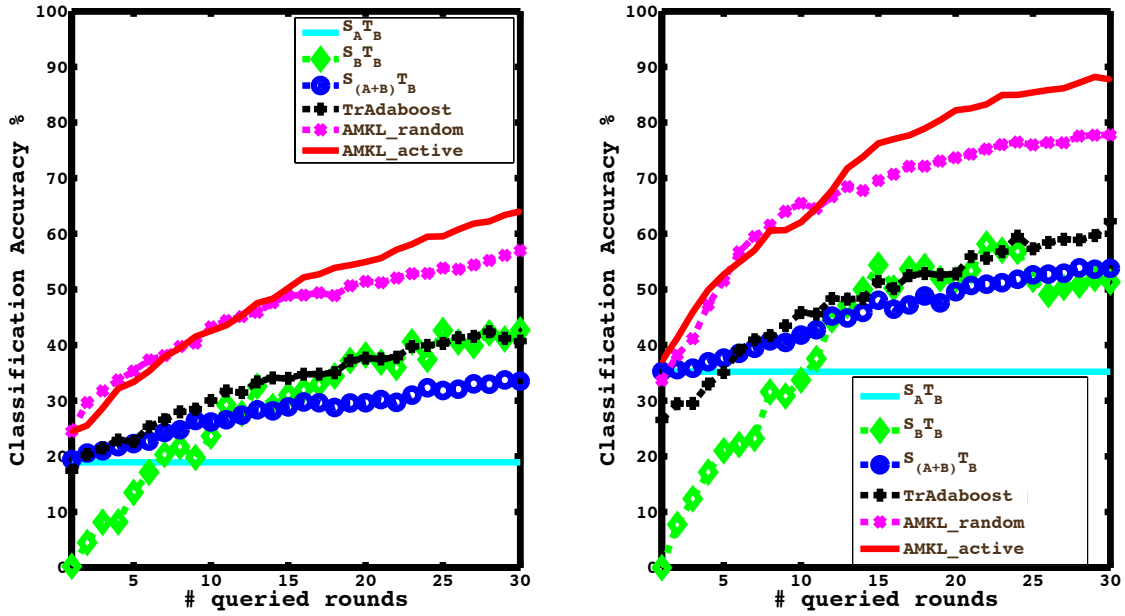


Figure 2.3: Classification accuracies with (left) single view and (right) 4 views.

ple/class) to label every round. To begin with, there are 100 unlabelled images per class in the target domain.

Fig.2.3 compares classification accuracies achieved using various approaches over 30 rounds of active learning. Evidently, we can see that our active transfer learning algorithm outperforms all the considered baselines. Clearly, our method efficiently learns about the target domain upon incorporating knowledge from a few target examples. Also, employing information from all four camera views achieves superior performance as compared to monocular analysis. Comparing AMKL_active with AMKL_random, we see that in both the monocular and multi-view cases, our approach outperforms AMKL_random after 10 rounds of AL, and the benefit of learning from the most informative samples is reflected in the fact that AMKL_active outperforms AMKL_random by more than 10% after 30 rounds while classifying with 4-view information.

Fig.2.4 shows the confusion matrix over 24 headpose classes using active transfer learning after 30 rounds. We can conclude that most of the target samples are correctly classified. Moreover, most of the misclassified samples belong

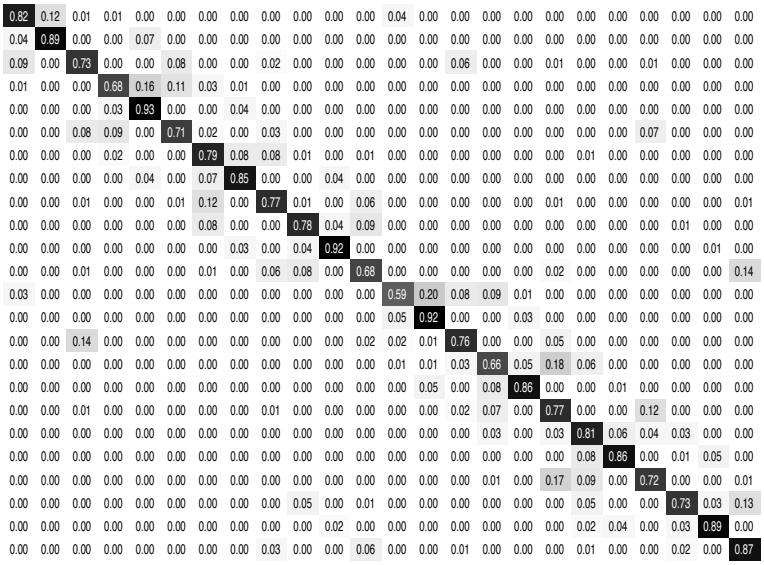


Figure 2.4: Confusion matrix over 24 classes for active DA after 30 rounds.

to nearby classes, which means the samples are only misclassified with respect to head tilt, while the head pan is classified correctly. This again demonstrates the robustness of our active DA framework for headpose estimation.

We also evaluate the effects of using five different types of loss function in the AL module– (i) Logistic loss $1/(1+e^{2yf(x)})$, (2) Exponential loss e^{-x}, (3) Hinge loss $(1-y)_+$, (4) Minimum margin loss e^{-100x} and (5) log loss $log(1/(1+x))$. Fig.2.5 presents the active transfer learning classification error obtained on these different loss function. We observe that hinge loss achieves the better performance among all loss functions, which implies that active transfer learning works optimally if identical loss functions are employed in the DA and AL modules.

Since querying sample labels for AL can also be done in a batch mode, we examine the extent of reduction in classification error for varying number of queried samples at every round. Fig.2.6 shows the progressive reduction in classification error with differing number of queried samples (4, 8 and 12 samples/class/round) for AL. From Fig.2.6, we can see that the classification accuracy is not influenced much by varying the number of queried samples per

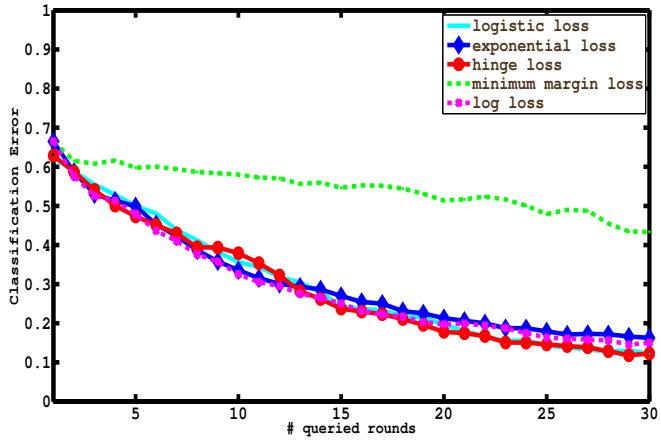


Figure 2.5: Evaluating active DA classification error with different loss functions.

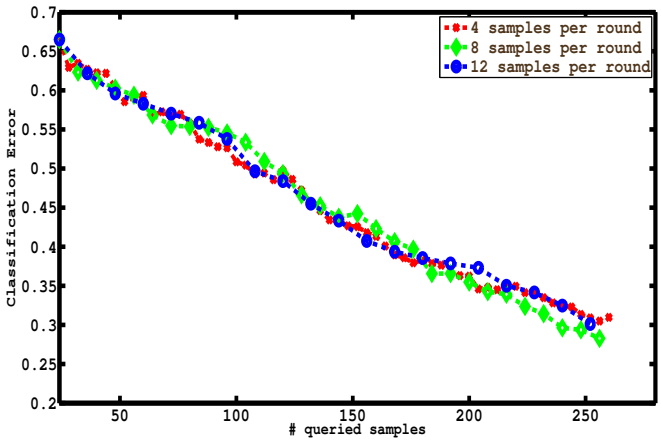


Figure 2.6: Evaluating our active DA framework with batch mode querying by varying number of queried samples/class/round.

round. However, choosing a moderate number of queried samples per round appears to be optimal since the error is minimal when 8 samples per round are queried as compared to querying 4 or 12 samples per round. Finally, we evaluate the robustness of our active DA framework to noisy labels. Fig.2.7 compares classification accuracies achieved with and without modeling for noisy labels in the AL module (steps 13–19 in Algorithm 1). Note that about 3% higher accuracy is achieved by accounting for noisy labels when using both monocular and 4-view image features.

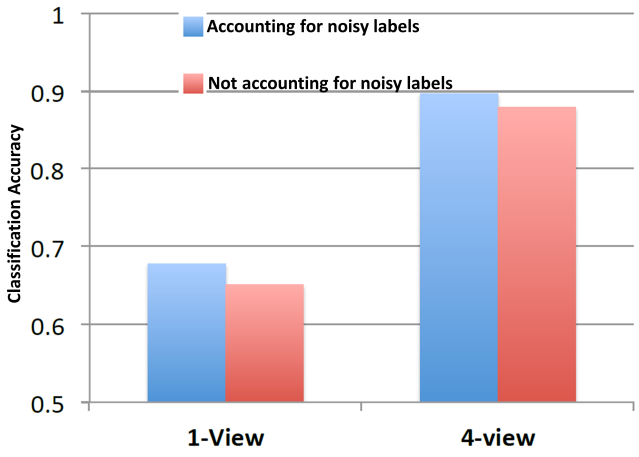


Figure 2.7: Evaluating active domain adaptation with noisy labels modeling strategy.

2.4.2 Cross-domain Berkeley Web Image Dataset

The Berkeley image dataset consists of three types of images: web images (from amazon), images from a digital SLR camera (high resolution image), and low-resolution webcam images, as shown in Fig.2.8. Each domain has 31 categories of images. While the digital SLR camera and webcam images capture the same objects, the viewpoint and image resolutions are different.

Our objective on the Berkeley dataset is to perform object recognition across image domains. For all the experiments, we report mean accuracy obtained on

Figure 2.8: Exemplars from the Berkeley web image dataset. (from top to bottom) Web (amazon), digital SLR camera (high resolution image) and webcam (low resolution image).

5 randomly selected train/test sets. SVM parameter $C = 1$ in all the experiments. For each object category, there are a small number of labelled samples in the target domain (3 in our experiment). For the source domain, we use 8 labels per category for *webcam/dslr* and 20 for *amazon*. As low-level visual descriptors, we use the pre-compute SURF features. A codebook of size 800 is constructed by k-means clustering. We firstly normalize the feature vector and then repeat the experiment as in [84]. Descriptions of the several baseline methods compared are as follows:

- $S_A T_B$ - We train on source domain A and test on target domain B.
- $S_B T_B$ - We train on target domain B and test on B.
- $S_{(A+B)} T_B$ - We train on both A and B, and test on B.
- [84] - A metric learning-based DA approach.
- $TrAdaboost$ ([23]) - DA based on the *Adaboost* algorithm.
- DA - DA with adaptive multiple kernel learning (AMKL) and randomly label target samples.

- ADA - DA and actively label target samples.
- ADAN - Proposed DA method accounting for noisy labels.

Tables 2.1, 2.2, 2.3 compare classification accuracies achieved with the different approaches when trained on images from the webcam, dslr and amazon domains respectively. We make the following observations from these tables: (i) Superior performance is always achieved using S_B as compared to S_A, which proves the need for DA for object recognition on the Berkeley dataset. (ii) While the inductive TrAdaboost and metric learning-based DA approaches perform favorably with respect to $S_{(A+B)}T_B$, they are generally outperformed by the AMKL-based DA approaches studied in this work. (iii) ADA outperforms DA considerably, implying that AL greatly benefits DA for object recognition. (iv) ADAN outperforms ADA by up to 5% on an average, implying that our approach which explicitly accounts for label noise greatly benefits AL. (iv) ADAN consistently produces the best recognition performance demonstrating the efficiency of the proposed active DA framework.

Commenting on the computational time required for our proposed algorithm, model training for cross-domain multi-view headpose estimation and object recognition required 20 minutes with cross-validation on a workstation with Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz $\times$ 17 processors implying that our algorithm can be applied on large-scale datasets.

2.5 Conclusion

In this chapter, we propose an active transfer learning framework which explicitly accounts for ambiguous labels provided by the domain expert. We also extend traditional active learning for binary classification to a multi-class setting through error-correcting output coding. Extensive experiments on cross-domain multi-view head-pose estimation and object recognition demonstrate the effectiveness of our proposed method. In particular, the ability to select the most

Table 2.1: Source domain - *webcam* images.

	webcam→dslr	webcam→amazon
$S_A T_B$	0.19 ± 0.02	0.09 ± 0.01
$S_B T_B$	0.37 ± 0.01	0.18 ± 0.02
$S_{(A+B)} T_B$	0.28 ± 0.02	0.15 ± 0.01
[84]	0.27 ± 0.02	0.19 ± 0.01
TrAdaboost ([23])	0.25 ± 0.02	0.17 ± 0.02
DA	0.35 ± 0.02	0.20 ± 0.01
ADA	0.61 ± 0.02	0.23 ± 0.01
ADAN	**0.65 ± 0.02**	**0.27 ± 0.02**

Table 2.2: Source domain - *dslr* images.

	dslr→webcam	dslr→amazon
$S_A T_B$	0.15 ± 0.01	0.04 ± 0.01
$S_B T_B$	0.40 ± 0.03	0.18 ± 0.02
$S_{(A+B)} T_B$	0.20 ± 0.02	0.08 ± 0.01
[84]	0.31 ± 0.03	0.15 ± 0.02
TrAdaboost ([23])	0.44 ± 0.03	0.10 ± 0.02
DA	0.49 ± 0.02	0.15 ± 0.02
ADA	0.59 ± 0.02	0.22 ± 0.02
ADAN	**0.63 ± 0.02**	**0.31 ± 0.02**

Table 2.3: Source domain - *amazon* images.

	amazon→dslr	amazon→webcam
$S_A T_B$	0.04 ± 0.02	0.08 ± 0.01
$S_B T_B$	0.36 ± 0.03	0.38 ± 0.02
$S_{(A+B)} T_B$	0.10 ± 0.03	0.14 ± 0.02
[84]	0.32 ± 0.02	0.48 ± 0.03
TrAdaboost ([23])	0.22 ± 0.03	0.38± 0.01
DA	0.28 ± 0.01	0.39 ± 0.02
ADA	0.36 ± 0.03	0.45 ± 0.01
ADAN	**0.40 ± 0.01**	**0.49 ± 0.03**

informative samples for active learning and handle label noise improves classification performance with respect to random sample selection.
In the next chapter, we will introduce another learning with shared information strategy, multi-task learning.

Chapter 3

Multi-task Dictionary Learning

In this chapter, we first propose a novel multi-task dictionary learning framework for the painting style recognition task. Then a novel supervised version of multi-task dictionary learning is proposed for image recognition.

3.1 Inferring Painting Style with Multi-Task Dictionary Learning[1]

Recent advances in imaging and multimedia technologies have paved the way for automatic analysis of visual art. Despite notable attempts, extracting relevant patterns from paintings is still a challenging task. Different painters, born in different periods and places, have been influenced by different schools of arts. However, each individual artist also has a unique signature, which is hard to detect with algorithms and objective features. In this chapter we propose a novel dictionary learning approach to automatically uncover the artistic style from paintings. Specifically, we present a multi-task learning algorithm to learn a style-specific dictionary representation. Intuitively, our approach, by automatically decoupling style-specific and artist-specific patterns, is expected to be more accurate for retrieval and recognition tasks than generic methods. To

[1]Gaowen Liu, Yan Yan, Elisa Ricci, Yi Yang, Yahong Han, Stefan Winkler, Nicu Sebe:Inferring Painting Style with Multi-Task Dictionary Learning. International Joint Conference on Artificial Intelligence (IJCAI): 2162-2168, 2015

demonstrate the effectiveness of our approach, we introduce the DART dataset, containing more than 1.5K images of paintings representative of different styles. Our extensive experimental evaluation shows that our approach significantly outperforms state-of-the-art methods.

3.1.1 Introduction

With the continuously growing amount of digitized art available on the web, classifying paintings into different categories, according to style, artist or based on the semantic contents, has become essential to properly manage huge collections. In addition, the widespread diffusion of mobile devices has led to an increased interest in the tourism industry for developing applications that automatically recognize the genre, the art movement, the artist, and the identity of paintings and provide relevant information to the visitors of museums.

Imaging and multimedia technologies have progressed substantially during the past decades, encouraging research on automatic analysis of visual art. Nowadays, art historians have even started to analyse art based on statistical techniques, *e.g.* for distinguishing authentic drawings from imitations [47]. However, despite notable attempts [14, 51, 57, 101], the automatic analysis of paintings is still a complex unsolved task, as it is influenced by many aspects, *i.e. low-level* features, such as color, texture, shading and stroke patterns, *mid-level* features, such as line styles, geometry and perspective, and *high-level* features, such as objects presence or scene composition.

In section 3.1 we investigate how to automatically infer the artistic style, *i.e.* *Baroque*, *Renaissance*, *Impressionism*, *Cubism*, *Postimpressionism* and *Modernism*, from paintings. According to Wikipedia, an artistic style is a "tendency with a specific common philosophy or goal, followed by a group of artists during a restricted period of time or, at least, with the heyday of the style defined within a number of years". Referring to paintings, the notion of *style* is more difficult to define than to perceive. Looking at Fig. 3.1, where images represen-

Figure 3.1: Given the images belonging to the *Baroque*, *Renaissance*, *Impressionism*, *Cubism*, *Postimpressionism*, *Modern* art movements, can you detect which ones correspond to the same style[1]?

tative of six art movements are shown, can you guess which ones belong to the same style? At the first glance, it may not be hard to group these images into different styles, *i.e.* (1) and (9), (4) and (8), even if you have never seen these paintings before. Indeed, human observers can easily match artworks from the same style and discriminate those originated from different art movements, even if no a-priori information is provided. That is because humans *recognize the style* by implicitly using both low-level cues such as lines or colors and more subtle compositional patterns.

[1]Answers: Cubism (1,9), Impressionism (2,7), Postimpressionism (3,10), Renaissance (4,8), Baroque (6,11), Modern (7,12).

Names and authors of paintings: 1) Bottle and Fishes, *Braque*; 2) Bouquet of Sunflowers, *Monet*; 3) Portrait of the Postman Joseph Roulin, *van Gogh*; 4) Christ Falling on the Way to Calvary, *Raphael*; 5) The Disintegration of the Persistence of Memory, *Dali*; 6) The Adoration of the Golden Calf, *Poussin*; 7) Portrait of Claude Renoir Painting, *Renoir*; 8) Death of Actaeon, *Titian*; 9) Bananas, *Gris*; 10) Vegetation

Recently, statistical methods have shown potential for supporting traditional approaches in the analysis of visual art by providing new, objective and quantifiable measures that assess the artistic style [14, 51, 101]. In this section we propose a dictionary learning approach for recognizing styles. Dictionary learning, which has proved to be highly effective in different computer vision and pattern recognition problems [31, 117], is a class of unsupervised methods for learning sets of over-complete bases to represent data efficiently. The aim of dictionary learning is to find a set of basis vectors such that an input vector can be represented as a linear combination of the basis vectors. In this section we propose a novel framework unifying multi-task and dictionary learning in order to simultaneously infer artist-specific and style-specific representations from a collection of paintings. Our intuition is that if we can build a style-specific dictionary representation by exploiting common patterns between artists of the same style with multi-task learning, more accurate results can be obtained for painting retrieval or recognition. For example, by automatically learning a dictionary for *Cubism* which captures the features associated to straight lines, we expect to easily detect that the paintings (1) and (9) in Fig.3.1 belong to the same category. Our experiments, conducted on the new DART (Dictionary ART) dataset, confirm our intuition and demonstrate that the learned dictionaries can be successfully used to recognize the artistic styles.

To summarize, the main contributions of section 3.1 are: (i) We are the first to introduce the idea of learning style-specific dictionaries for automatic analysis of paintings. (ii) A novel multi-task dictionary learning approach is proposed through embedding all tasks into an optimal learned subspace. Our multi-task learning strategy permits to effectively separate artist-specific and style-specific patterns, improving recognition performances. The proposed machine learning framework is a generic one and can be easily applied to other problems. (iii) We collected the DART dataset which contains paintings from different art

Tropicale, Martinique, *Gauguin*; 11) The Night Watch, *Rembrandt*; 12) Living Still Life, *Dali*.

movements and different artists.

3.1.2 Related Work

3.1.2.1 Automatic Analysis of Paintings

In literature, [22] were the first to borrow ideas from classification systems for automatic analysis of visual art and studied the differences between paintings and photographs. Image features such as edges, spatial variation of colors, number of unique colors, and pixel saturation were used for classification. [57] compared van Gogh with his contemporaries by statistical analysis of a massive set of automatically extracted brushstrokes. [14] introduced the problem of artistic image annotation and retrieval and proposed several solutions using graph-based learning techniques. [101] proposed a SOM-based model for studying and visualizing the relationships among painting collections of different painters. [123] presented an analysis of the affective cues extracted from abstract paintings by looking at low-level features and employing a bag-of-visual-words approach. Few works focused specifically on inferring style from paintings [51, 86]. However, none of these works have studied the problem of decoupling artist-specific and style-specific patterns as we do with our multi-task dictionary learning framework.

3.1.2.2 Dictionary and Multi-task Learning

Dictionary learning has been shown to be able to find succinct representations of stimuli. Recently, it has been successfully applied to a variety of problems in computer vision, pattern recognition and image processing, *e.g.* image classification [117], denoising [31]. Different optimization algorithms [1, 55] have also been proposed to solve dictionary learning problems. However, as far as we know, there is no research work on learning dictionary representations for recognizing artistic styles.

Multi-task learning [6, 113, 114] methods aim to simultaneously learn clas-

sification and regression models for a set of related tasks. This is typically advantageous as compared to considering single tasks separately and not exploiting their relationships. The goal of multi-task learning is to improve the performance by learning models for multiple tasks jointly. This works particularly well if these tasks have some commonality while are all slightly undersampled. However, there is hardly any work on combining multi-task and dictionary learning problems. [82] developed an efficient online algorithm for dictionary learning from multiple consecutive tasks based on the K-SVD algorithm. Another notable exception is [70] where theoretical bounds are provided to study the generalization error of multi-task dictionary learning algorithms. [19, 20, 122] proposed different convex formulations for feature selection problems. These works are very different from ours, since we focus on a specific applicative scenario and propose a novel multi-task dictionary learning algorithm.

3.1.3 Learning Style-specific Dictionaries

In this section we present our multi-task dictionary learning approach for inferring style-specific representations from paintings. In the following we first describe the chosen feature descriptors and then the proposed learning algorithm.

3.1.3.1 Feature Extraction from Paintings

Color, composition and brushstrokes are considered to be the three most important components in paintings. Therefore, to represent each painting, we construct a 37-dimensional feature vector as proposed in [101], including color, composition and lines informations (Fig.3.2).

Color. Following [101], the color features are computed as a function of luminance and hue. They are: (i) The visual temperature of color (the feel of warmth or coldness of color), as the wavelengths of the visible color light waves are con-

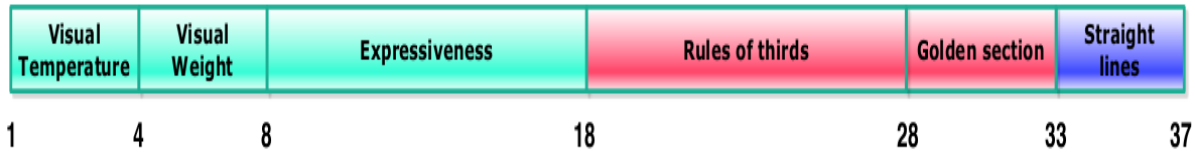


Figure 3.2: Extracted features: color (light blue), composition (red) and lines (blue).

sidered to be related to the human perception of color temperatures. Different emotions can be expressed by using cold or warm color temperatures. (ii) The visual weight of color (the feel of heaviness of color). From the perspective of psychology, people usually feel that a darker color is heavier and a lighter color is lighter. (iii) The expressiveness of color (the degree of contrast including the contrast between luminance, saturation, hue, color temperature, and color weight). Global and local contrast features are both used to measure the differences between pixel and image regions.

Composition. The composition represents the spatial organization of visual elements in a painting. For each image we compute a saliency map. The saliency map is divided into three parts both horizontally and vertically and we consider the mean salience for each of the nine sections to compute the "rule of thirds". Additionally, properties of the most salient region such as size, elongation, rectangularity and the most salient point are used to represent properties of 'golden section' composition principles. In details, elongation measures the symmetricity along the principal axes, rectangularity measures how close it is to its minimum bounding rectangle, the most salient point is the global maximum of the saliency map.

Lines. Lines in paintings are generally perceived as edges. Different styles of paintings or different painters may favour a certain type of line. To interpret the concepts of lines, the Hough Transform is adopted to find straight lines that are above a certain threshold (longer than 10 pixels). The mean slope, mean length, and standard deviation of slopes of all the detected straight lines are calculated.

Algorithm 2: Learning artist-specific and style-specific dictionaries.

Input:
Samples $\mathbf{X}_1, ..., \mathbf{X}_k$ from K tasks
Subspace dimensionality s, dictionary size l, regularization parameters λ_1, λ_2.
Output:
Optimized $\mathbf{P} \in \mathbb{R}^{d \times s}$, $\mathbf{C}_k \in \mathbb{R}^{n_k \times l}$, $\mathbf{D}_k \in \mathbb{R}^{l \times d}$, $\mathbf{D} \in \mathbb{R}^{l \times s}$.
Initialize $\mathbf{P}$ using any orthonormal matrix
Initialize $\mathbf{C}_k$ with l_2 normalized columns
repeat
  Compute $\mathbf{D}$ using Algorithm 2 in [66]
  for $k = 1 : K$
    Compute $\mathbf{D}_k$ using Algorithm 2 in [66]
    Compute $\mathbf{C}_k$ using FISTA [8]
  end for
  Compute $\mathbf{P}$ by eigendecomposition of $\mathbf{B} = \mathbf{X}'(\mathbf{I} - \mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}')\mathbf{X}$;
until *Convergence*;

3.1.3.2 Multi-task Dictionary Learning

Intuitively, in this and in many other applications [52, 67], it is reasonable to expect that more accurate recognition results are achieved if class specific dictionaries are adopted rather than generic ones. To this end, in this section we demonstrate that better classification performance are obtained when we consider a style-specific dictionary for each artistic style. In details, we propose to jointly learn a set of artist-specific dictionaries and discover the underlying style-specific dictionary projecting data in a low dimensional subspace.

More formally, for each painting style we consider K tasks and the k-th task corresponds to the k-th artist. Each task consists of data samples denoted by $\mathbf{X}_k = [\mathbf{x}_k^1, \mathbf{x}_k^2, ..., \mathbf{x}_k^{n_k}]$, $\mathbf{X}_k \in \boldsymbol{R}^{n_k \times d}, k = 1, ..., K$, where $\mathbf{x}_k^i \in \boldsymbol{R}^d$ is a d-dimensional feature vector and n_k is the number of samples in the k-th task. We propose to learn a shared subspace across all tasks, obtained by an orthonormal projection $\mathbf{P} \in \boldsymbol{R}^{d \times s}$, where s is the dimensionality of the subspace. In this learned subspace, the data distribution from all tasks should be similar to each other. Therefore, we can code all tasks together in the shared subspace and achieve better coding quality. The benefits of this strategy are: (i) We can improve each

individual coding quality by transferring knowledge across all tasks. (ii) We can discover the relationship among different tasks (artists) via coding analysis. (iii) The common dictionary among tasks, *i.e.* the style-specific dictionary, can be learned by embedding all tasks into a good sharing subspace. These objectives can be realized solving the optimization problem:

$$\begin{aligned}\min_{\mathbf{D}_k,\mathbf{C}_k,\mathbf{P},\mathbf{D}} \sum_{k=1}^{K} \|\mathbf{X}_k - \mathbf{C}_k\mathbf{D}_k\|_F^2 + \lambda_1 \sum_{k=1}^{K} \|\mathbf{C}_k\|_1 \\ +\lambda_2 \sum_{k=1}^{K} \|\mathbf{X}_k\mathbf{P} - \mathbf{C}_k\mathbf{D}\|_F^2\end{aligned} \tag{3.1}$$

$$\text{s.t.} \quad \begin{cases} \mathbf{P}'\mathbf{P} = \mathbf{I} \\ (\mathbf{D}_k)_{j\cdot}(\mathbf{D}_k)'_{j\cdot} \leq 1, \quad \forall j = 1, ..., l \\ \mathbf{D}_{j\cdot}\mathbf{D}'_{j\cdot} \leq 1, \quad \forall j = 1, ..., l \end{cases}$$

where $\mathbf{D}_k \in \boldsymbol{R}^{l\times d}$ is an over-complete (artist-specific) dictionary ($l > d$) with l prototypes of the k-th task, $(\mathbf{D}_k)_{j\cdot}$ in the constraints denotes the j-th row of $\mathbf{D}_k$, and $\mathbf{C}_k \in \boldsymbol{R}^{n_k\times l}$ corresponds to the sparse representation coefficients of $\mathbf{X}_k$. In the third term of Eq.3.1, $\mathbf{X}_k$ is projected by $\mathbf{P}$ into the subspace to explore the relationship among different tasks. $\mathbf{D} \in \boldsymbol{R}^{l\times s}$ is the (style-specific) dictionary learned in the tasks-shared subspace and $\mathbf{D}_{j\cdot}$ in the constraints denotes the j-th row of $\mathbf{D}$. Moreover, $\mathbf{I}$ is the identity matrix, $(\cdot)'$ denotes the transpose operator and λ_1 and λ_2 are regularization parameters. The first constraint guarantees the learned $\mathbf{P}$ to be orthonormal, and the second and third constraints prevent the learned dictionary to be arbitrarily large. In our objective function, we learn a dictionary $\mathbf{D}_k$ for each task k and one shared dictionary $\mathbf{D}$ among k tasks. When $\lambda_2 = 0$, Eq.3.1 reduces to the traditional dictionary learning on separated tasks. It is fundamental to underline the difference between $\mathbf{D}$ and $\mathbf{D}_k$: $\mathbf{D}$ is the learned style-specific dictionary and $\mathbf{D}_k$ is the dictionary associated the k-th artist in each style. In Eq.3.1, we share the same coefficient $\mathbf{C}_k$ in the global and in the task-specific reconstruction error terms. This is actually meant to enforce

the coherence between artist-specific and style-specific dictionaries found in the low dimensional subspace.

3.1.3.3 Optimization

To solve the problem in Eq.3.1, we adopt an alternating optimization algorithm. The proposed algorithm is summarized in Algorithm 2. The source code for the optimization will be made available online. In details, we optimize with respect to $\mathbf{D}$, $\mathbf{D}_k$, $\mathbf{C}_k$ and $\mathbf{P}$ respectively in four steps as follows:

Step 1: Fixing $\mathbf{D}_k$, $\mathbf{C}_k$, P, compute D. Considering the matrices $\mathbf{X} = [\mathbf{X}'_1, ..., \mathbf{X}'_k]'$, $\mathbf{C} = [\mathbf{C}'_1, ..., \mathbf{C}'_k]'$, we obtain $\sum_{k=1}^{K} ||\mathbf{X}_k\mathbf{P} - \mathbf{C}_k\mathbf{D}||_F^2 = ||\mathbf{XP} - \mathbf{CD}||_F^2$. Therefore Eq.3.1 is equivalent to:

$$\begin{aligned} \min_{\mathbf{D}} \quad & ||\mathbf{XP} - \mathbf{CD}||_F^2 \\ \text{s.t.} \quad & \mathbf{D}_{j\cdot}\mathbf{D}'_{j\cdot} \leq 1, \quad \forall j = 1, ..., l \end{aligned}$$

This is equivalent to the dictionary update stage in traditional dictionary learning algorithms. We adopt the dictionary update strategy of Algorithm 2 in [66] to efficiently solve it.

Step 2: Fixing D, $\mathbf{C}_k$, P, compute $\mathbf{D}_k$. To compute $\mathbf{D}_k$ we solve:

$$\begin{aligned} \min_{\mathbf{D}_k} \quad & ||\mathbf{X}_k - \mathbf{C}_k\mathbf{D}_k||_F^2 \\ \text{s.t.} \quad & (\mathbf{D}_k)_{j\cdot}(\mathbf{D}_k)'_{j\cdot} \leq 1, \quad \forall j = 1, ..., l \end{aligned} \tag{3.2}$$

Similarly to Step 1, solving (3.2) corresponds to the update stage for dictionary learning in case of k tasks. Then, to compute $\mathbf{D}_k$ we also use the approach described in Algorithm 2 in [66].

Step 3: Fixing $\mathbf{D}_k$, P, D, compute $\mathbf{C}_k$. Eq.3.1 is equivalent to:

$$\begin{aligned} \min_{\mathbf{C}_k} \sum_{k=1}^{K} ||\mathbf{X}_k - \mathbf{C}_k\mathbf{D}_k||_F^2 + \lambda_1 \sum_{k=1}^{K} ||\mathbf{C}_k||_1 \\ +\lambda_2 \sum_{k=1}^{K} ||\mathbf{X}_k\mathbf{P} - \mathbf{C}_k\mathbf{D}||_F^2 \end{aligned}$$

This problem can be decoupled into $n' = n_1 + n_2 + ... + n_k$ distinct problems:

$$\min_{\mathbf{c}_k^i} \|\mathbf{x}_k^i - \mathbf{c}_k^i \mathbf{D}_k\|_2^2 + \lambda_1 \|\mathbf{c}_k^i\|_1 + \lambda_2 \|\mathbf{x}_k^i \mathbf{P} - \mathbf{c}_k^i \mathbf{D}\|_2^2 \tag{3.3}$$

We adopt the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [8] to solve the problems in Eq.3.3. FISTA solves the optimization problems in the form of $\min_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) + r(\boldsymbol{\mu})$, where $f(\boldsymbol{\mu})$ is convex and smooth, and $r(\boldsymbol{\mu})$ is convex but non-smooth. We adopt FISTA since it is a popular tool for solving many convex smooth/non-smooth problems and its effectiveness has been verified in many applications. In our setting, we denote the smooth term part as $f(\mathbf{c}_k^i) = \|\mathbf{x}_k^i - \mathbf{c}_k^i \mathbf{D}_k\|_2^2 + \lambda_2 \|\mathbf{x}_k^i \mathbf{P} - \mathbf{c}_k^i \mathbf{D}\|_2^2$ and the non-smooth term part as $g(\mathbf{c}_k^i) = \lambda_1 \|\mathbf{c}_k^i\|_1$.

Step 4: Fixing $\mathbf{D}_k$, $\mathbf{C}_k$, D, compute P. Considering $\mathbf{X} = [\mathbf{X}_1', ..., \mathbf{X}_k']'$, $\mathbf{C} = [\mathbf{C}_1', ..., \mathbf{C}_k']'$, we solve:

$$\begin{aligned} \min_{\mathbf{P}} \quad & \|\mathbf{XP} - \mathbf{CD}\|_F^2 \\ \text{s.t.} \quad & \mathbf{P'P} = \mathbf{I} \end{aligned} \tag{3.4}$$

Substituting $\mathbf{D} = (\mathbf{C'C})^{-1}\mathbf{C'XP}$ back into the above function, we obtain:

$$\begin{aligned} \min_{\mathbf{P}} \quad & \text{tr}(\mathbf{P'X'}(\mathbf{I} - \mathbf{C}(\mathbf{C'C})^{-1}\mathbf{C'})\mathbf{XP}) \\ \text{s.t.} \quad & \mathbf{P'P} = \mathbf{I} \end{aligned}$$

The optimal **P** is composed of eigenvectors of the matrix $\mathbf{B} = \mathbf{X'}(\mathbf{I} - \mathbf{C}(\mathbf{C'C})^{-1}\mathbf{C'})\mathbf{X}$ corresponding to the s smallest eigenvalues.

After the optimized dictionaries are obtained for styles and artists, the final classification of a test image is based on computing its sparse coefficient and calculating the minimal reconstruction error, similarly to [67, 119].

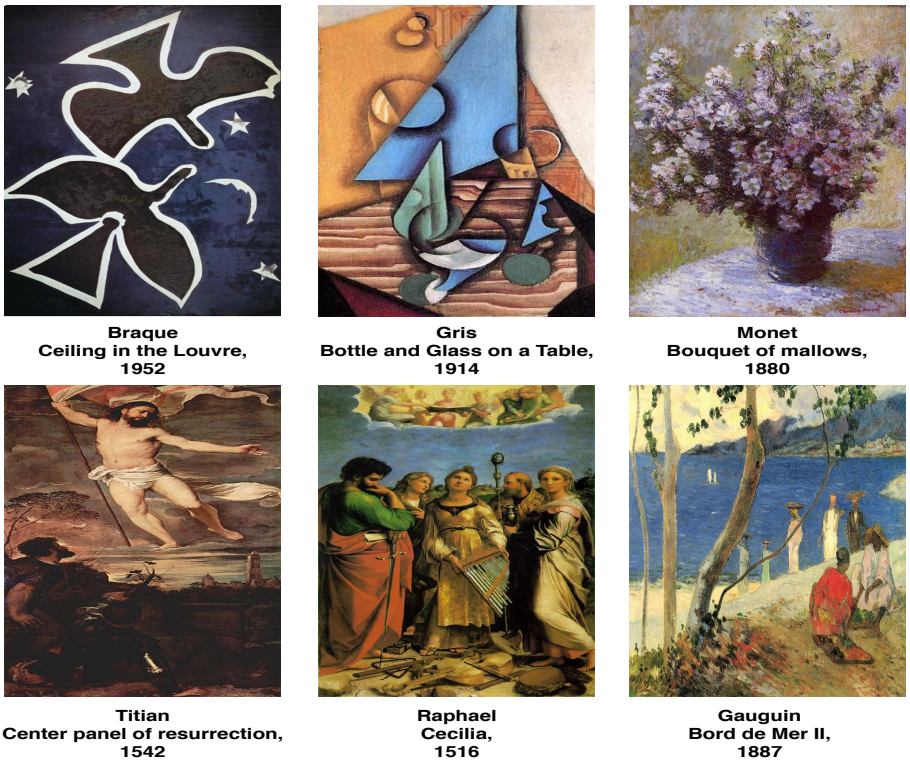


Figure 3.3: Examples of paintings from the DART dataset. Each image is associated with a detailed description containing year, artist and painting name.

3.1.4 Experimental Results

In this section we introduce the DART dataset and evaluate the effectiveness of our method.

3.1.4.1 Dataset

The DART dataset contains paintings collected from the web representing six different artistic styles, *i.e.*, *Baroque*, *Cubism*, *Impressionism*, *Postimpressionism*, *Renaissance* and *Modern*. Examples with a detailed description for artists, painting name and year as recorded in DART are shown in Fig.3.3. For each style the painting of at least three artists have been collected. As shown in Table 3.1, there are totally 1616 paintings in the DART dataset. There is a high variability in paintings as each artist typically developed different painting techniques and styles as time passed. Therefore, for each painter, we ensured that

Table 3.1: Structure of the DART dataset.

Artistic Style	Artists	# of paintings
Baroque	Rubens	60
	Rembrandt	104
	Poussin	117
Cubism	Braque	113
	Gris	119
	Picasso	62
Impressionism	Monet	108
	Renoir	109
	Manet	58
Post-impressionism	van Gogh	134
	Gauguin	136
	Odilon	69
Renaissance	Raphael	67
	Titian	92
	Bosch	46
	Caravaggio	59
Modern	Mondrian	60
	Frida	45
	Dali	58

the selected artworks cover a wide range of techniques and subjects. We also ensured that the paintings are from different periods of the artist life. To the best of our knowledge, DART is the largest high quality art dataset available with paintings and associated descriptions so far.

3.1.5 Experimental Setup and Baselines

In our experiments we randomly split the dataset into two parts, half for training and half for testing. We repeated the experiments ten times. The average results and associated standard deviations are reported. We set the regularization parameters, the subspace dimensionality s and the dictionary size l with cross-validation.

We compare the proposed method with several state-of-the-art single-task

dictionary learning and multi-task learning methods. Specifically we consider (1) *Support Vector Machine* (SVM); (2) *Elastic Net* (EN), as it is the classifier used for painting style analysis in [51]; (3) *Dictionary Learning by Aggregating Tasks* (AT-DL), *i.e.* performing single task dictionary learning by simply aggregating data from all tasks; (4) *Locality-constrained Linear Coding* (LLC) [99], a method which uses the locality constraints to project each descriptor into its local-coordinate system and integrates the projected coordinates by max pooling to generate the final representation; (5) *Graph Structure Multi-Task Learning*[3] (GSMTL) [133], a state-of-the-art multi-task learning method imposing graph structure to exploit tasks relationship; (6) *Dirty Model Multi-Task Learning*[3] (DMMTL) [49], a multi-task learning algorithm based on ℓ_1/ℓ_q-norm regularization; (7) *Robust Multi-Task learning*[3] (RMTL) [21], a multi-task learning approach which imposes a low rank structure capturing task-relatedness and detects outlier tasks.

3.1.5.1 Quantitative Evaluation

We conduct extensive experiments to evaluate the effectiveness of the proposed method in recognizing artistic styles. Table 3.2 compares our approach with different single-task dictionary learning and multi-task methods. From Table 3.2, the following observations can be made: (i) Our proposed style-specific dictionary learning method significantly outperforms generic single task methods such as SVM and EN. (ii) Multi-task learning approaches (GSMTL, DMMTL, RMTL) always perform better than single-task dictionary learning (AT-DL, LLC) since they consider the correlation among paintings of different artists with the same style. (iii) Our approach performs better than the other multi-task learning methods, due to its unique ability of combining multi-task and dictionary learning. By introducing style-specific dictionaries a more discriminative data representation is obtained.

[3] http://www.public.asu.edu/~jye02/Software/MALSAR/

Table 3.2: Comparison with baseline methods.

Methods	Average accuracy
SVM	0.564 ± 0.004
EN [51]	0.624 ± 0.007
AT-DL	0.595 ± 0.003
LLC [99]	0.642 ± 0.003
GSMTL [133]	0.681 ± 0.010
DMMTL [49]	0.651 ± 0.005
RMTL [21]	0.672 ± 0.006
Ours	**0.745 ± 0.003**

Table 3.3: Evaluation on different features combinations.

Features	Average accuracy
Raw Pixels	0.527 ± 0.004
Color	0.533 ± 0.002
Composition	0.571 ± 0.008
Lines	0.489 ± 0.003
Color + Composition	0.632 ± 0.005
Color + Lines	0.598 ± 0.004
Composition + Lines	0.675 ± 0.006
Color + Composition + Lines	**0.745 ± 0.005**

Fig. 3.4(left) shows the confusion matrix obtained with the proposed method. *Cubism* achieves relative high recognition accuracies compared with other styles, which is reasonable since the paintings belonging to *Cubism* contain many "long lines" compared with other styles. This aspect is evident observing Fig. 3.1. Moreover, many *Impressionism* and *Postimpressionism* paintings are misclassified into the other class because these styles are more correlated. In the literature, *Postimpressionism* was influenced by *Impressionism*. Indeed, *Postimpressionism* was meant to extend *Impressionism*. The painters continued to use vivid colors and brushtrokes and focused on real-life subjects, but they were more interested to emphasize geometric forms, use unnatural colors and distort the original forms for more expressive effects.

We also evaluate our approach with respect to different parameters,

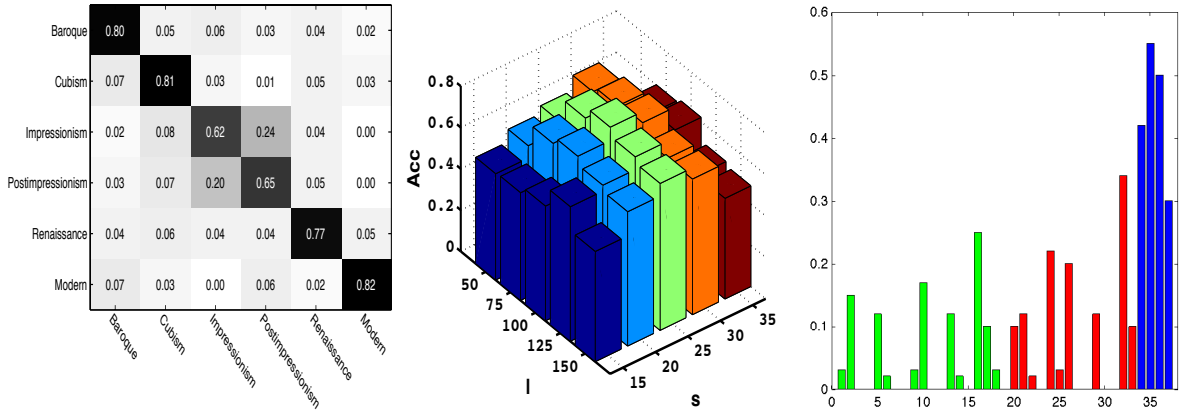


Figure 3.4: (Left) Confusion Matrix on DART dataset. (Middle) Performance at varying dictionary size l and subspace dimensionality s. (Right) Visualization of contributions of each component for the *Cubism* style. Different colors represent different components, *i.e.* color (green), composition (red) and lines (blue).

namely the dictionary size l and the different subspace dimensionality s. Fig. 3.4(middle) shows that the proposed method achieves the best results when the dictionary size is 100 and the subspace dimensionality is 25. Too large or too small values for dictionary size and subspace dimensionality tend to decrease the performance. We also analyze the convergence of the proposed approach.

It is also interesting to investigate the contributions of each component (color, composition, and lines) for painting style classification. To evaluate this, we set the dictionary length equal to the dimensions of the feature vector and averaged the learned sparse codes for each style. Fig. 3.4(right) visualizes the contribution of each component for the *Cubism* style. We observe that the line features contribute the most to the recognition of the *Cubism* style. We also quantitatively evaluate the importance of different features on recognizing all styles as shown in Table 3.3. Raw pixels, color, composition, lines and their combinations are considered. Experimental results shows that using high-level features is advantageous with respect to simply using raw pixels. Moreover, combining all the heterogeneous features is greatly beneficial in terms of accuracy. While raw pixels are not appropriate for classification, to give a better idea

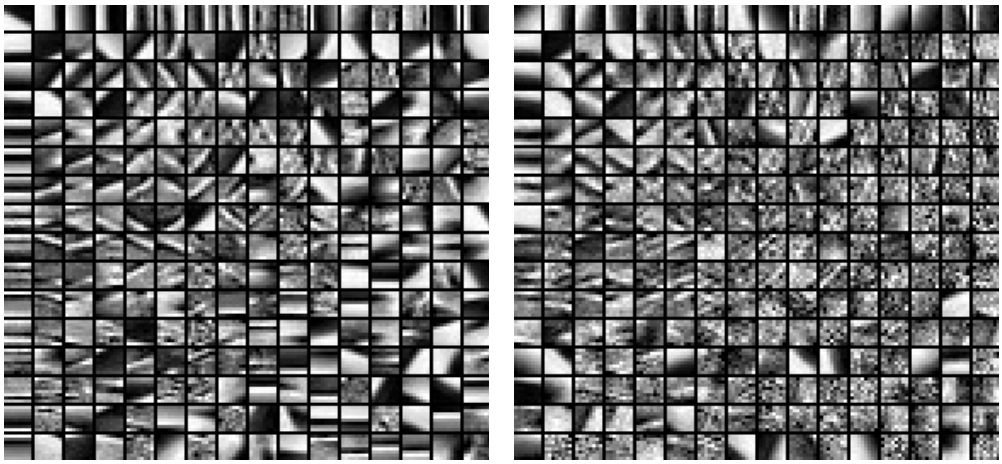


Figure 3.5: Visualization of learned dictionaries when using raw pixels features for (left) *Cubism* and (right) *Renaissance*.

of the output of our method, we use pixel values as features to learn the dictionary for each specific style. Fig. 3.5 visualizes the qualitative learned dictionaries for the *Cubism* and the *Renaissance* style, respectively. It is interesting to notice that the learned dictionaries share some similarity while many visual patterns are different. This clearly implies the necessity of learning style-specific dictionaries for paintings classification.

Finally, to further validate the proposed feature representation, we show a phylogenetic tree reflecting the similarities among artists (Fig. 3.6). The similarities are measured by euclidean distance among the average values of our feature vectors. Then a hierarchical clustering algorithm is applied. We can clearly see that painting collections of the same artistic styles are much more similar to each other than painting collections of different art movements (*e.g.* Dali is clustered with Frida Kahlo and Mondrian rather than with Rubens or Picasso).

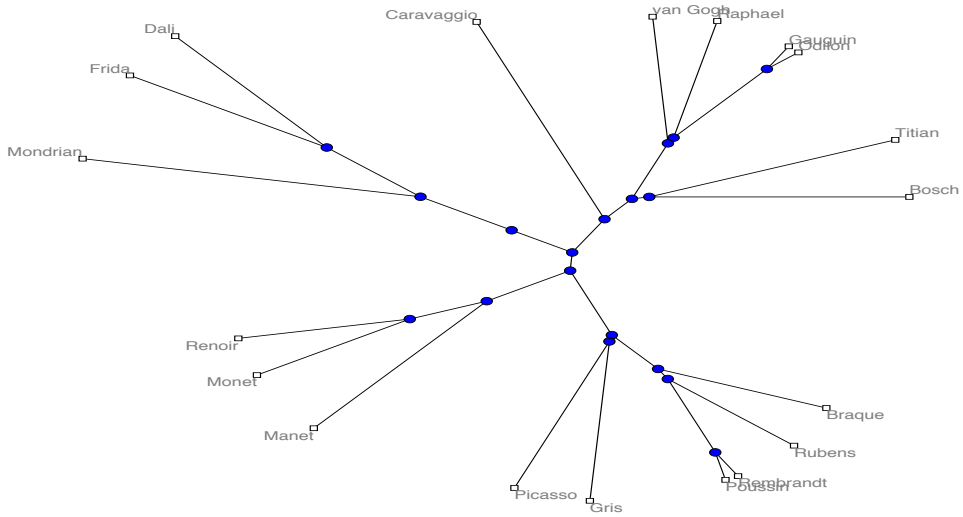


Figure 3.6: The phylogenetic tree reflecting the similarities among artists. (Figure is best viewed under zoom).

3.2 Discriminative multi-task dictionary learning for image classification[2]

Following the work on painting style recognition with multi-task dictionary learning in section 3.1, we step further to present our work on image classification with dictionary learning in section 3.2.

Sparse coding (Dictionary learning) was shown to be able to find succinct representations of stimuli. Recently, it has been successfully applied to a variety of problems in image processing analysis. Sparse coding models data vectors as a linear combination of a few elements from a dictionary. However, most existing sparse coding methods are applied for a single task on a single dataset. The learned dictionary is then possibly biased towards the specific dataset and lacks of generalization abilities. In light of this, in section 3.2 we propose a multi-task sparse coding approach by uncovering a shared subspace among heterogeneous datasets. The proposed multi-task coding strategy leveraged the

[2]Gaowen Liu, Yan Yan, Jingkuan Song, Nicu Sebe: Minimizing dataset bias: Discriminative multi-task sparse coding through shared subspace learning for image classification. International Conference on Image Processing (ICIP): 2869-2873, 2014

commonality benefit from different datasets. Moreover, our multi-task coding framework is capable of direct classification by incorporating label information. Experimental results show that the dictionary learned by our approach has more generalization abilities and our model performs better classification compared to the model learned from only one dataset or the model learned from simply pooling different datasets together.

3.2.1 Introduction

Recently, sparse coding has been successfully applied to a variety of problems in image processing, *e.g.* image classification [117], image denoising [31] and image segmentation [68]. Different optimization algorithms [1, 55] have also been proposed to solve sparse coding problems. Sparse coding was shown to be able to find succinct representations of stimuli and model data vectors as a linear combination of a few elements from a dictionary. However, in terms of image recognition task, most sparse coding methods work on a single task on a single dataset. Therefore, the learned dictionary can be highly biased towards the specific dataset and can be lack of generalization abilities. The dataset selection bias [25, 90, 91, 97, 127] is a common problem in research. In [97], the authors point out that despite the best efforts of the dataset creator, the datasets always appear to have strong build-in bias. Nowadays, most experimental evaluations are often done within a heterogeneous dataset, so it is questionable that the results are a reliable indicator of true generalization. The learned dictionary based on a dataset could not probably carry enough general information to be applied to different datasets.

Multi-task learning [4] aims to simultaneously learn classification/regression models for a set of related tasks. This typically leads to better models as compared to a learner that does not consider task relationships. Multi-task learning has been successfully applied to different computer vision and image processing problems, such as image classification [126], human headpose estimation

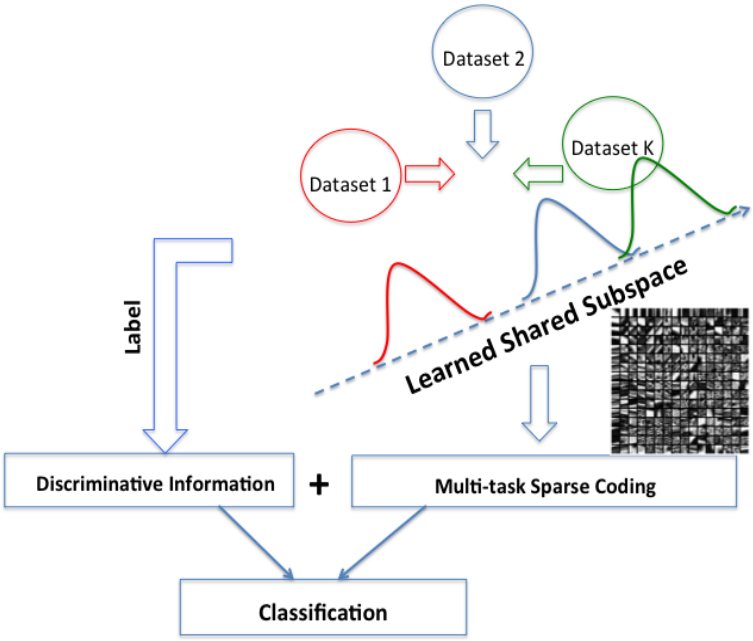


Figure 3.7: Framework of Discriminative Multi-task Sparse Coding through Shared Subspace Learning.

[107, 110] and human action recognition [104, 109]. In [70], authors provide theoretical bounds on the generalization error of sparse coding for multi-task learning and transfer learning. Considering the goal of minimizing the dataset bias and dictionary generalization, in this section we propose a new multi-task sparse coding approach by uncovering a shared subspace of different datasets. Moreover, since the traditional sparse coding framework cannot directly classify the data, we incorporate label information into our sparse coding framework. This enables our proposed model to be directly used for classification. The overall framework of our proposed method is shown in Fig.1. Different tasks are embedded in the shared subspace to be coded together and then the label information from different tasks is incorporated. The codes learned by this framework contain discriminative information and this can be used for classification. Experimental results show that the dictionary learned by our approach has more generalization abilities and our model has better classification performance compared to the model learned from only a single dataset or the model

learned from simply pooling different datasets together.

To sum up, section 3.2 makes the following contributions:

- We propose a multi-task sparse coding algorithm by exploiting shared subspace learning and an efficient solver is devised;
- The learned dictionary based on our approach has more generalization abilities towards different datasets;
- The relationship among different tasks could be discovered through the learned shared subspace;
- Our model can be directly used for classification due to the adding of label information.

The rest of this section is organized as follows. Section 3.2.2 elaborates the details of our Muti-task Sparse Coding approach and illustrates how we design it directly for classification. Section 3.2.3 reports the experimental results, followed by conclusion in section 3.2.4.

3.2.2 Problem Formulation

In this section, we first present our multi-task sparse coding framework and then illustrate how we make the framework suitable for classification. An optimization algorithm is further proposed to solve the problem.

3.2.2.1 Multi-task Sparse Coding

Suppose we are given K tasks and each task consists of data samples denoted by $X_k = \{x_k^1, x_k^2, ..., x_k^{n_k}\} \in R^{n_k \times d}, (k = 1, ..., K)$, where $x_k^i \in R^d$ is a d-dimensional feature vector, n_k is the number of samples in the k-th task. We are going to learn a shared subspace, obtained by an orthonormal projection $W \in R^{d \times s}$ across all tasks, where s is the dimensionality of the subspace. In this learned subspace, the data distribution from all tasks should be similar to each other. Therefore, we can code all tasks together in the shared subspace and achieve better coding quality. The benefits of this strategy are: (i) we can improve each individual

coding quality by transferring knowledge across all tasks. (ii) we can discover the relationship among different datasets via coding analysis. We consider the following optimization problem:

$$\begin{aligned} \min_{D_k,C_k,W,D} & \sum_{k=1}^{K} \|X_k - C_k D_k\|_F^2 + \lambda_1 \sum_{k=1}^{K} \|C_k\|_1 \\ & +\lambda_2 \sum_{k=1}^{K} \|X_k W - C_k D\|_F^2 \\ s.t. \quad & \begin{cases} W^T W = I \\ (D_k)_{j\cdot}(D_k)_{j\cdot}^T \leq 1, \quad \forall j = 1, ..., l \\ D_{j\cdot} D_{j\cdot}^T \leq 1, \quad \forall j = 1, ..., l \end{cases} \end{aligned} \tag{3.5}$$

where $D_k \in R^{l \times d}$ is an overcomplete dictionary ($l > d$) with l prototypes of k-th task, $(D_k)_{j\cdot}$ in the constraints denotes the j-th row of D_k. $C_k \in R^{n_k \times l}$ are the sparse representation coefficients of X_k. In the third term of Eqn.1, X_k is projected by W to the subspace to explore the relationship among different tasks. $D \in R^{l \times s}$ is the dictionary learned in the subspace. $D_{j\cdot}$ in the constraints denotes the j-th row of D and I is an identity matrix. $(\cdot)^T$ denotes the transpose operator. λ_1 and λ_2 are regularization parameters. The first constraint guarantees the learned W to be orthonormal. The second and third constraints prevent learned dictionary being arbitrarily large. When $\lambda_2 = 0$, Eqn.1 converts to the traditional sparse coding on separated tasks.

3.2.2.2 Discriminative Multi-task Sparse Coding for Classification

It is well-known that the traditional sparse coding framework is not suitable for classification and the learned dictionary is merely used for signal reconstruction. To circumvent this problem, researchers have developed several algorithms to learn a classification-oriented dictionary in a supervised learning fashion by exploring the label information. In this subsection, we extend our proposed Multi-task Sparse Coding of Eqn.1 to be suitable for classification.

Assuming the k-th task has m_k classes, the label information of the k-th task is $Y_k = \{y_k^1, y_k^2, ..., y_k^{n_k}\} \in R^{n_k \times m_k}, (k = 1, ..., K)$, $y_k^i = [0, ..., 0, 1, 0, ..., 0]$ (the position of a non-zero element indicates the class). $\Theta_k \in R^{l \times m_k}$ is the parameter of the k-th task classifier. Inspired by [130], we consider the following optimization

Algorithm 3: Algorithm for Multi-task Sparse Coding.

Input:

K tasks Data $(X_1, ..., X_k)$ and Label $(Y_1, ..., Y_k)$;

Subspace dimensionality s, Dictionary size l, Regularization parameters λ_1, λ_2, λ_3.

Output:

Optimized $W \in \mathbb{R}^{d\times s}$, $C_k \in \mathbb{R}^{n_k\times l}$, $D_k \in \mathbb{R}^{l\times d}$, $D \in \mathbb{R}^{l\times s}$, $\Theta_k \in \mathbb{R}^{l\times m_k}$.

1: Initialize W using any orthonormal matrix;
2: Initialize C_k with l_2 normalized columns;
3: **repeat**
Compute D using Algorithm 2 in [66];
for $k = 1 : K$
Compute D_k using Algorithm 2 in [66];
Adopting FISTA [8] to solve C_k;
$\Theta_k = (C_k^T C_k)^{-1} C_k^T Y_k$;
end for
Compute W by eigen decomposition of $X^T(I - C(C^T C)^{-1}C^T)X$;
until *Convergence*;

problem:

$$\begin{aligned}\min_{D_k,C_k,\Theta_k,W,D} &\sum_{k=1}^{K} \|X_k - C_k D_k\|_F^2 + \lambda_1 \sum_{k=1}^{K} \|C_k\|_1 \\ &+\lambda_2 \sum_{k=1}^{K} \|X_k W - C_k D\|_F^2 + \lambda_3 \sum_{k=1}^{K} \|Y_k - C_k \Theta_k\|_F^2 \\ s.t. \quad &\begin{cases} W^T W = I \\ (D_k)_{j\cdot}(D_k)_{j\cdot}^T \le 1, \quad \forall j = 1, ..., l \\ D_{j\cdot} D_{j\cdot}^T \le 1, \quad \forall j = 1, ..., l \end{cases}\end{aligned} \tag{3.6}$$

Compared with Eqn.1, we added the last term in Eqn.2 to incorporate the discriminative power for classification . This objective function can simultaneously achieve a desired dictionary with good representation power and support optimal discrimination of the classes for multi-task setting. To solve the proposed objective problem of Eqn.2, we adopt the alternating minimization algorithm to optimize it with respect to D, D_k, C_k, Θ_k and W respectively. We propose Algorithm 3 to solve the objective function of Eqn.2.

After the optimized Θ is obtained, the final classification of a test image is based on its sparse coefficient c_k^i, which carries the discriminative information. We can simply apply the linear classifier $c_k^i \Theta_k$ to obtain the predicted label of

the image.

3.2.3 Experiments

In this section, we conduct the experiments to evaluate the effectiveness of our framework.

3.2.3.1 Datasets

Since our aim is to evaluate the generalization ablity of our proposed method, we use the Animals with Attributes (AwA) dataset[1] and a subset (25 animal classes) of the Caltech-101 dataset[2]. In this way, images from two datasets are all animals which are reasonable for multi-task learning to share similarity among tasks. The examples from these two datasets are shown in Fig.3.8. We can observe that the Caltech-101 dataset has small spatial variances (the target object is often in the central part of the images). However, the AwA dataset has large spatial variances. This dataset selection bias gives us the possibility to evaluate our proposed method.

3.2.3.2 Experiment Settings

For both datasets, SIFT features are extracted from 16×16 patches with a stride of 6 pixels. Spatial pyramid features based on SIFT features are extracted for 3-level (1×1, 2×2, 4×4) pyramid. The codebook size for spatial pyramid is set as $512 \times (1 + 4 + 16)$. In each spatial sub-region of the spatial pyramid, the vector quantization codes are pooled together using max pooling to form the pooled feature. The final spatial pyramid feature is reduced to 1000 dimensions by PCA. We set the regularization parameters in the range of $\{10^{-2}, 10^{-1}, ..., 10^{2}\}$. The subspace dimensionality s is set by searching the grid from {100, 200, 400, 600}. For the experiments in this section, we tried four

[1] http://attributes.kyb.tuebingen.mpg.de/

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

Figure 3.8: Example images from AwA dataset (top) and Caltech-101 dataset (bottom).

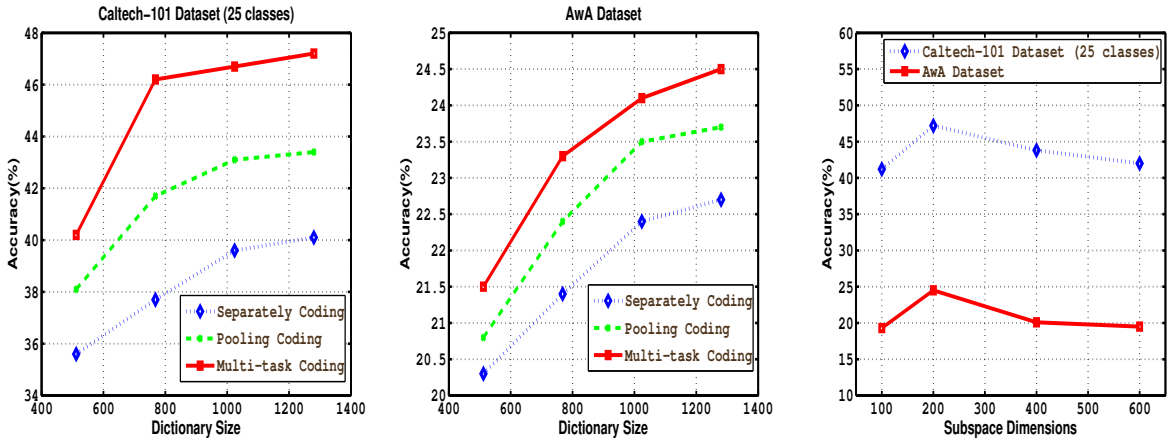
different dictionary sizes from {512, 768, 1024, 1280}. The presented results denote the mean classification accuracy corresponding to five randomly chosen training sets.

3.2.3.3 Quantitative Evaluation

In this subsection, we report the quantitative evaluation results. We compare our multi-task sparse coding to the following baselines:

- Separately coding: Performing sparse coding separately on each dataset;
- Pooling coding: Performing sparse coding simply putting all datasets together.

Fig.3.9(a)-(b) show experimental results when we vary the dictionary size (5 training samples per class and 200 dimensional subspace are used). We can observe that (i) Coding different datasets together performs better than coding each dataset separately. This proves that the dataset bias exists. (ii) Our multi-task sparse coding strategy always outperforms single task sparse coding or simply

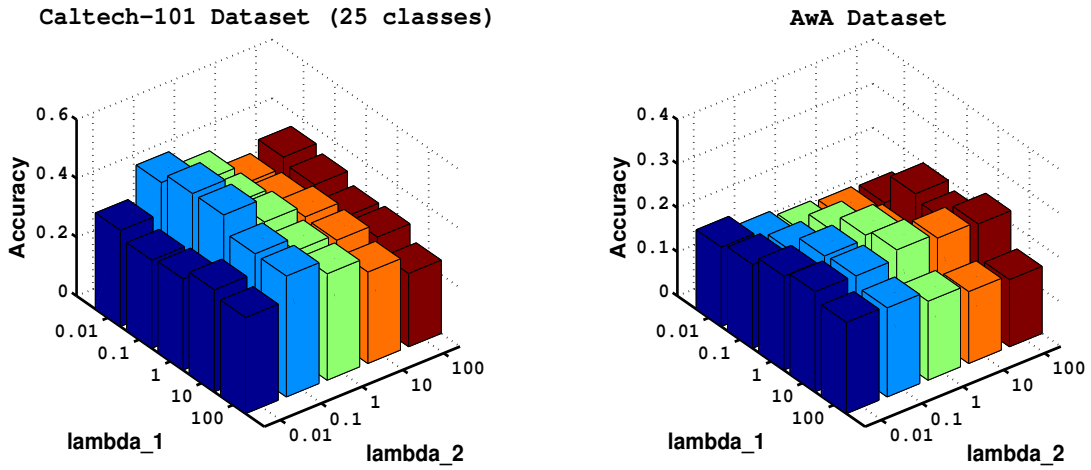


(a) (b) (c)

Figure 3.9: Performance comparisions (5 training samples per class used) (a) Different dictionary size on Caltech-101 dataset; (b) Different dictionary size on AwA dataset; (c) Different subspace size on Caltech-101 and AwA dataset.

coding datasets together with various dictionary sizes. This shows that sharing information among multiple tasks could be used for minimizing the dataset selection bias. Fig.3.9(c) shows experimental results with various subspace dimensions (5 training samples per class and dictionary length of 1024 are used). We observe that the best performance is achieved when the subspace dimensionality is 200 and the performance is sensitive to the subspace dimensionality, which means that we need to embed different tasks into a good subspace to share information among different tasks effectively.

We also study the parameter sensitivity of the proposed method in Fig.3.10. Here, we fix $\lambda_3 = 1$ (discriminative information contribution fixed) and analyze the regularization parameters λ_1 and λ_2. We observe that the proposed method is more sensitive to λ_2 compared with λ_1, which demonstrates the importance of the subspace for multi-task sparse coding. At last, we compare our proposed method to several sparse coding based [117, 130] and subspace feature selection based [64] classification approaches. [117] is a spatial-pyramid image representation based on sparse codes and max pooling. [130] is a dictionary-learning approach by adding a discriminative term into the objective function



(a) (b)

Figure 3.10: Sensitivity study of parameters λ_1 and λ_2 (5 training samples per class used) on (a) Caltech-101 dataset; (b) AwA dataset.

of the original K-SVD algorithm. [64] is a method which combined shared feature subspace uncovering and joint feature selection with sparsity. However, all these methods are only in the single task settings. The comparision results are listed in Table 1. From the table, we observe that our proposed multi-task coding achieves higher classification accuracy compared to the methods based on single task, which shows the effectiveness of our proposed method.

Table 3.4: Recognition accuracy (5 training samples per class)

	Datasets	
Method	Caltech-101 (25 classes)	AwA
Proposed	**0.492**	**0.265**
Ma [64]	0.433	0.203
Zhang [130]	0.451	0.211
Yang [117]	0.454	0.209

3.2.4 Conclusion

In section 3.1, we investigated how to automatically infer painting styles from the perspective of dictionary learning and we proposed a novel multi-task dictionary learning approach to discover a low dimensional subspace where a style-specific dictionary representation can be computed. We conducted extensive experiments to evaluate our algorithm on the new DART dataset. Our results show that our style-specific approach performs significantly better than a generic one and that the proposed multi-task method achieves higher accuracy than state of the art dictionary learning algorithms.
In section 3.2, we have proposed a supervised version of multi-task dictionary learning based on different datasets. The proposed model learns a shared subspace to transfer knowledge among different datasets. The model is also able to perform classification and we apply it for image annotation. Experimental results show that the dictionary learned by our approach has more generalization abilities and our model performs better classification compared to the model learned from only one dataset or the model learned from simply pooling different datasets together.

In next chapter, we will introduce an unsupervised multi-task clustering framework for first-person vision activity recognition.

Chapter 4

Activity recognition via Multi-task Clustering[1]

Recognizing human activities from videos is a fundamental research problem in computer vision. Recently, there has been a growing interest in analysing human behaviour from data collected with wearable cameras. First-person cameras continuously record several hours of their wearers' life. To cope with this vast amount of unlabelled and heterogeneous data, novel algorithmic solutions are required. In this chapter, we propose a multi-task clustering framework for address the problem of activity of daily living analysis from visual data gathered from wearable cameras. Our intuition is that, even if the data are not annotated, it is possible to exploit the fact that the tasks of recognizing everyday activities of multiple individuals are related, since typically people perform the same actions in similar environments (*e.g.* people working in an office often read and write documents). In our framework, rather than clustering data from different users separately, we propose to look for clustering partitions which are coherent among related tasks. Specifically, two novel multi-task clustering algorithms, derived from a common optimization problem, are introduced. Our experimental evaluation, conducted both on synthetic data and on publicly available first-

[1] Yan Yan, Elisa Ricci, Gaowen Liu, Nicu Sebe: Egocentric Daily Activity Recognition via Multitask Clustering. IEEE Transactions on Image Processing 24(10): 2984-2995, 2015

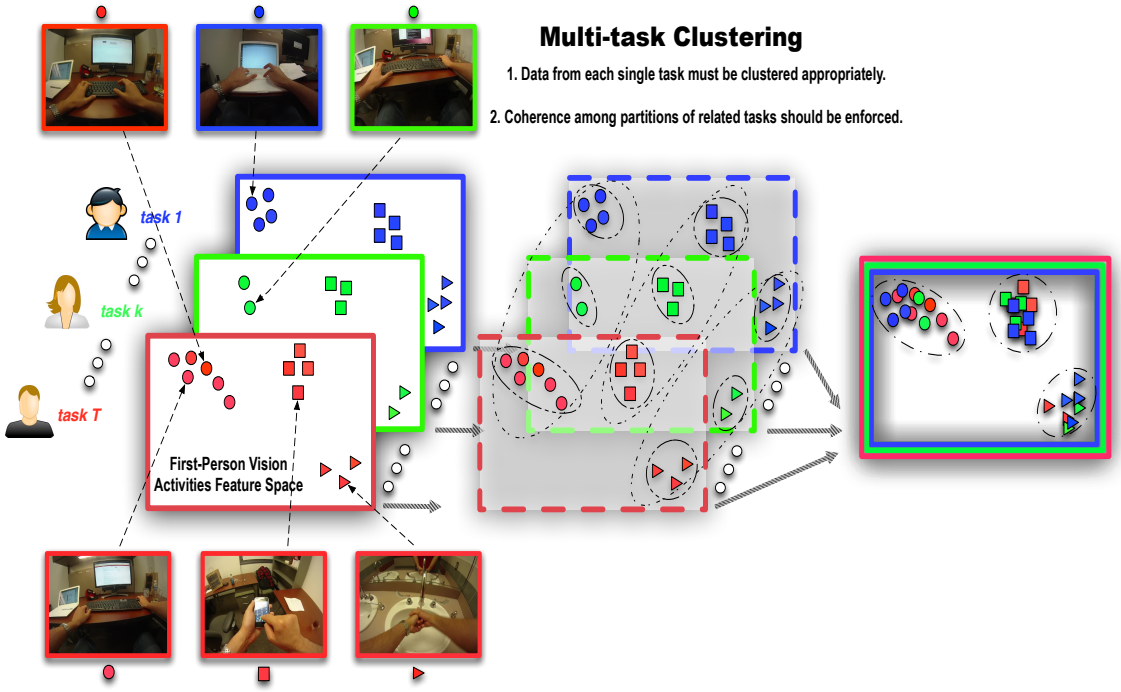


Figure 4.1: Overview of our multi-task clustering approach for FPV activity recognition (Figure best viewed in color).

person vision datasets, shows that the proposed approach outperforms several single task and multi-task learning methods.

4.1 Introduction

Research in wearable sensor-based activity recognition leverages the data automatically collected from sensors embedded into mobile devices to predict the user daily activities in real-time. RFID, GPS and accelerometers represent the most popular wearable sensors and several works [16, 94] have already proposed to exploit them for inferring people behaviors. Nowadays, wearable cameras are becoming increasingly common among consumers. Wearable cameras can be employed in many different applications, such as life-logging, ambient assisted living, personal security and drivers' assistance. It is intuitive that, while GPS and inertial sensors may suffice for detecting simple activities (*i.e.*

running, walking), only by analyzing visual informations from wearable cameras more complex behaviors can be inferred.

Activity of Daily Living (ADL) analysis has attracted considerable attention in the computer vision and image processing communities [7, 56, 71, 80]. Analyzing visual streams recorded from video surveillance cameras to automatically understand *what people do* is a challenging task [98]. It implies not only to infer the activities of a single individual, but also to recognize the environment where he/she operates, the people with whom he/she interacts, the objects he/she manipulates and even his/her future intentions. While much progress has been made in this area, recent works [50] have demonstrated that the traditional "third-person" view perspective (*i.e.* employing fixed cameras monitoring the user environment) may be insufficient for recognizing user activities and intentions and that wearable cameras provide a valid alternative.
In this chapter, we consider the problem of ADL analysis from a first-person vision (FPV) perspective. Among the many challenges arising in this context, one particular issue is related to the fact that wearable cameras are intended to record the entire life of a person. Thus, a huge amount of visual data is automatically generated. Moreover, labels are usually not available since the annotation would require a massive human effort. As a consequence, algorithms which are both scalable and able to operate in an unsupervised setting are required. To face these challenges, we propose to cast the problem of egocentric daily activity recognition within a Multi-Task Learning (MTL) framework. When considering the tasks of inferring everyday activities of several individuals, it is natural to assume that these tasks are related. For example, people working in an office environment typically perform the same activities (*e.g.* working in front of a personal computer, reading and writing documents). Similarly, people at home in the morning usually make breakfast and brush their teeth. In this chapter we argue that, when performing activity recognition, learning from data of several targets simultaneously is advantageous with respect to consid-

ering each person separately. For example, if there are limited data for a single person, typical clustering methods may fail to discover the correct clusters and leveraging auxiliary sources of information (*e.g.* data from other people) may improve the performance. However, simply combining data from different people together and applying a traditional clustering approach does not necessarily increase accuracy, because the data distributions of single tasks can be different (*i.e.* visual data corresponding to different people may exhibit different features). To address these problems, we propose a novel Multi-Task Clustering (MTC) framework from which we derive two different algorithms. Our approach ensures that the data of each single task are clustered appropriately and simultaneously enforces the coherence between clustering results of related tasks (Fig. 4.1). To demonstrate the validity of our method we first conduct experiments on synthetic data and compare it with state-of-the-art single task and multi-task learning algorithms. Then, we show that our approach is effective in recognizing activities in an egocentric setting and we consider two recent FPV datasets, the FPV activity of daily living dataset [77] and the coupled ego-motion and eye-motion dataset introduced in [74]. This chapter extends our previous work [105].

To summarize, the contributions of this chapter are the following: (i) To our knowledge, this is the first work proposing a multi-task clustering framework for FPV activity recognition. Most papers on MTL for human activity analysis [65, 108] focus on video collected from fixed cameras and mostly rely on supervised methods. (ii) Our work is one of the few works presenting an *unsupervised* approach for MTL. The proposed MTC methods are novel and two efficient algorithms are derived for solving the associated optimization problems. (iii) Our learning framework is general and many other computer vision and pattern recognition applications can benefit from using it.

The chapter is organized as follows. Section 4.2 reviews related work on first person vision activity recognition and supervised/unsupervised multi-task

learning. In Section 4.3 our MTC framework for FPV activity recognition is described in details. The experimental results are reported in Section 4.4. We then conclude in Section 4.5.

4.2 Related Work

In this section, we review prior works in (i) FPV activity analysis, (ii) supervised MTL and (iii) multi-task clustering.

4.2.1 First-person Vision Activity Analysis

Automatically analyzing human behavior from videos is a widely researched topic. Many previous works have focused on recognizing everyday activities [56, 71, 80]. In [71] features based on the velocity history of tracked keypoints are proposed for detecting complex activities performed in a kitchen. A kitchen scenario is also analyzed by Rohrbach *et al.* [80] and an approach for fine-grained activity recognition is presented. More recently, RGB-D sensors are exploited for ADL analysis [56] and improved performance is obtained with respect to approaches based on traditional cameras. However, all these works consider a "third-person" view perspective, *i.e.* they are specifically designed to analyse video streams from fixed cameras.

The challenges of inferring human behavior from data collected by wearable cameras are addressed in [34–36, 74, 75, 77, 95]. Aghazadeh *et al.* [75] proposed an approach for discovering anomalous events from videos captured from a small camera attached to a person's chest. In [62] a video summarization method targeted to FPV is presented. Fathi *et al.* [36] introduced a method for individuating social interactions in first-person videos collected during social events. Some recent works have focused on FPV-ADL analysis considering different scenarios (*e.g.* kitchen, office, home) [34, 35, 74, 77, 95]. In [77] Pirsi *et al.* introduced some features based on the output of multiple object detectors.

In [74] the task of recognizing egocentric activities in an office environment is considered and motion descriptors extracted from an outside looking camera are combined with features describing the user eye movements captured by an inside looking camera. In [95] activity recognition in a kitchen scenario (*i.e.* multiple subjects preparing different recipes) is considered. A codebook learning framework is proposed in order to alleviate the problem of the large within-class data variability due to the different execution styles and speed among different subjects. Ryoo *et al.* [83] investigated multi-channel kernels to integrate global and local motion information and presented a new activity recognition methodology that explicitly models the temporal structures of FPV data. In [78] an approach for temporal segmentation of egocentric videos into twelve hierarchical classes is presented. Differently from these previous works, in this chapter we address the problem of FPV ADL analysis proposing a multi-task learning framework.

4.2.2 Supervised Multi-task Learning

Multi-task learning methods [15] have recently proved to be particularly effective in many applications, such as complex event detection [111], object detection [85], head pose estimation [112], image classification [63], painting style recognition [61], etc. The idea of MTL is simple: given a set of related tasks, by simultaneously learning the corresponding classification or regression models, improved performance can be achieved. Usually, the advantages of MTL over traditional approaches based on learning independent models are particularly pronounced when the number of samples in each task is limited.

To capture the tasks dependencies a common approach is to constrain all the learned models to share a common set of features. This motivates the introduction of a group sparsity term, *i.e.* the ℓ_1/ℓ_2-norm regularizer as in [5]. This approach works well in ideal cases. However, in practical applications, simply using a ℓ_1/ℓ_2-norm regularizer may not be effective since not every task is re-

lated to all the others. To this end, the MTL algorithm based on the dirty model is proposed in [49] with the aim to identify irrelevant (outlier) tasks. Similarly, robust multi-task learning is introduced in [21]. In some cases, the tasks exhibit a sophisticated group structure and it is desirable that the models of tasks in the same group are more similar to each other than to those from a different group. To model complex task dependencies several clustered multi-task learning methods have been introduced [48, 131, 132]. In computer vision, MTL have been previously proposed in the context of visual-based activity recognition from fixed cameras and in a supervised setting [65, 108, 125]. In this chapter, we consider the more challenging FPV scenario where no annotated data are provided.

4.2.3 Multi-task Clustering

Many works on MTL focused on a supervised setting. Only few [40, 54, 129] have considered the more challenging scenario where the data are unlabelled and the aim is to predict the cluster labels in each single task. Gu *et al.* [40] presented an algorithm where a shared subspace is learned for all the tasks. Zhang *et al.* [129] introduced a MTC approach based on a pairwise agreement term which encourages coherence among clustering results of multiple tasks. In [54] the k-means algorithm is revised from a Bayesian nonparametric viewpoint and extended to MTL. None of these works have focused on the problem of visual-based activity recognition.

In this chapter, we propose two novel approaches for multi-task clustering. The first one is inspired by the work in [129] but it is based on another objective function and thus on a radically different optimization algorithm. Furthermore, in the considered application, it provides superior accuracy with respect to [129]. Our second approach instead permits to easily integrate prior knowledge about the tasks and the data of each task (*e.g.* temporal consistency among subsequent video clips). Moreover, it relies on a convex optimization problem,

thus avoids the issues related to local minima of previous methods [40, 54, 129].

4.3 Multi-task Clustering for First-person Vision Activity Recognition

In this section, we first introduce the motivation behind our approach, together with an overview of the proposed framework. Then, two different MTC algorithms, namely Earth Mover's Distance Multi-Task Clustering (EMD-MTC) and Convex Multi-task Clustering (CMTC), and their application to the problem of FPV ADL recognition are described.

4.3.1 Motivation and Overview

We consider the videos collected from wearable cameras of several people performing daily activities. No annotation is provided. We only assume that people perform about the same tasks, a very reasonable assumption in the context of ADL analysis.

To discover people activities, we consider T related tasks corresponding to T different people[1] and we introduce a MTC approach. For each task (person) t, a set of samples $X^t = \{\mathbf{x}_1^t, \mathbf{x}_2^t, ..., \mathbf{x}_{N_t}^t\}$ is available, where $\mathbf{x}_j^t \in \boldsymbol{R}^d$ is the d-dimensional feature vector describing the j-th video clip and N_t is the total number of samples associated to the t-th task. We want to segment the entire video clip corresponding to user t into parts, *i.e.* we want the data in the set X^t to be grouped into K_t clusters. Furthermore, as we assume the tasks to be related, we also require that the resulting partitions are consistent with each other. This is a reasonable assumption in the context of everyday activity recognition where people perform about the same activities. Note that the number of required partitions K_t can be different for different tasks, as different people can perform slightly different types of activities. Our assumptions are verified in

[1] This is not a constraint. In this chapter we focus on detecting activities for each user by exploiting related information from other users.

the context of ADL recognition. For example, typical activities in the morning are preparing breakfast, eating and brushing teeth. Therefore, when analysing video streams collected by wearable cameras of different users, it is reasonable to expect that the recordings will capture the same or at least very similar activities. To automatically discover these activities, we formulate the following optimization problem corresponding to multi-task clustering:

$$\min_{\boldsymbol{\Theta}_t} \sum_{t=1}^{T} \Lambda(X^t, \boldsymbol{\Theta}^t) + \lambda \sum_{t=1}^{T} \sum_{s=t+1}^{T} \Omega(\boldsymbol{\Theta}^t, \boldsymbol{\Theta}^s) \tag{4.1}$$

The term $\Lambda(\cdot)$ represents a reconstruction error which must be minimized by learning the optimal task-specific model parameters $\boldsymbol{\Theta}^t$ (*i.e.* typically the cluster centroids and the associated assignment matrix), while $\Omega(\cdot)$ is an "agreement" term imposing that, since the multiple tasks are related, also the associated model parameters should be similar. Under this framework, in this chapter we propose two different algorithms for MTC. To stress the generality of our framework, we apply the proposed algorithms in two different FPV scenarios: an office environment where people are involved in typical activities such as browsing the web or writing documents and a home environment where a chest mounted camera records users' activities such as opening a fridge or preparing tea. To perform experiments we use two publicly available datasets, corresponding to the scenarios described above: the FPV office dataset introduced in [74] and the FPV ADL dataset described in [77]. Both datasets contain visual streams recorded from an outside-looking wearable camera while the office dataset also has information about eye movements acquired by an inside-looking camera. In the following subsections we describe the proposed MTC algorithms and the adopted feature descriptors.

Notation: In the following $\mathbf{A}_{i.}$, $\mathbf{A}_{.j}$ denote respectively the i-th row and the j-th column of the matrix $\mathbf{A}$. We also denote with $(\cdot)'$ the transpose operator, $N = \sum_{t=1}^{T} N_t$ is the total number of datapoints, while $\mathbf{X} \in \boldsymbol{R}^{N \times d}$,

$\mathbf{X} = [\mathbf{X}^{1\prime}\ \mathbf{X}^{2\prime}\ \ldots\ \mathbf{X}^{T\prime}]'$ is the data matrix obtained by concatenating the task specific matrices $\mathbf{X}^t \in \boldsymbol{R}^{N_t \times d}$, $\mathbf{X}^t = [\mathbf{x}_1^t\ \mathbf{x}_2^t\ ...\ \mathbf{x}_{N_t}^t]'$.

4.3.2 Earth Mover's Distance Multi-task Clustering

Given the task data matrices $\mathbf{X}^t$, we are interested in finding the centroid matrices $\mathbf{C}^t \in \boldsymbol{R}^{K_t \times d}$, and the cluster indicators matrices $\mathbf{W}^t \in \boldsymbol{R}^{N_t \times K_t}$ by solving the following optimization problem:

$$\min_{\mathbf{C}^t, \mathbf{W}^t} \sum_{t=1}^{T} \left\| \mathbf{X}^t - \mathbf{W}^t \mathbf{C}^t \right\|_F^2 + \lambda \sum_{t=1}^{T} \sum_{s=t+1}^{T} \Omega_E(\mathbf{C}^t, \mathbf{W}^t, \mathbf{C}^s, \mathbf{W}^s)$$

The first term in the objective function is a relaxation of the traditional k-means objective function for T separated data sources. The agreement term $\Omega_E(\cdot)$ is added to explore the relationships between clusters of different data sources and it is defined as follows:

$$\begin{aligned}
\Omega_E(\mathbf{C}^t, \mathbf{W}^t, \mathbf{C}^s, \mathbf{W}^s) = \quad & \min_{f_{ij}^{st} \geq 0} \quad \sum_{i=1}^{K_t} \sum_{j=1}^{K_s} f_{ij}^{st} (\mathbf{C}_{i.}^t - \mathbf{C}_{j.}^s)' (\mathbf{C}_{i.}^t - \mathbf{C}_{j.}^s) \\
& \text{s.t.} \quad \sum_{j=1}^{K_s} f_{ij}^{st} = \sum_{n=1}^{N_t} \mathbf{W}_{ni}^t \quad \forall t, i \\
& \qquad \sum_{i=1}^{K_t} f_{ij}^{st} = \sum_{n=1}^{N_s} \mathbf{W}_{nj}^s \quad \forall s, j \\
& \qquad \sum_{i=1}^{K_t} \sum_{j=1}^{K_s} f_{ij}^{st} = 1 \quad \forall s, t
\end{aligned}$$

It consists in the popular Earth Mover's Distance (EMD) [81] computed considering the signatures $\mathcal{T}$ and $\mathcal{S}$ obtained by clustering the data associated to task t and s separately, *i.e.* $\mathcal{T} = \{(\mathbf{C}_1^t., w_t^1),\ \ldots,\ (\mathbf{C}_{K_t}^t., w_t^{K_t})\}$, $w_t^i = \sum_{n=1}^{N_t} \mathbf{W}_{ni}^t$, and $\mathcal{S} = \{(\mathbf{C}_1^s., w_s^1),\ \ldots,\ (\mathbf{C}_{K_s}^s., w_s^{K_s})\}$, $w_s^i = \sum_{n=1}^{N_s} \mathbf{W}_{ni}^s$. In practice $\mathbf{C}_i^t.$ and $\mathbf{C}_j^s.$ are the cluster centroids and w_i^s, w_i^t denote the weights associated to each cluster (approximating the number of datapoints in each cluster). In practice $\Omega_E(\cdot)$ rep-

Algorithm 4: Algorithm for solving (4.2).

Input: the data matrices $\mathbf{X}^1, \mathbf{X}^2$, the numbers of clusters K_1, K_2, the parameter λ.

1: Initialize $\mathbf{F}$ as an identity matrix.
2: Initialize $\mathbf{W} > 0$ with l_1 normalized columns and $\mathbf{P} > 0$ with l_1 normalized rows.
3: **repeat**

Given $\mathbf{W}^k$, $\mathbf{P}^k$, compute $\mathbf{F}^{k+1}$ solving (4.4).
Given $\mathbf{F}^{k+1}$, $\mathbf{P}^k$, compute: $\mathbf{W}^{k+1} = \max(0, \mathbf{W}^k - \alpha_k \nabla_{\mathbf{W}} \Delta(\mathbf{P}^k, \mathbf{W}^k, \mathbf{F}^{k+1}))$.
Given $\mathbf{F}^{k+1}$, $\mathbf{W}^{k+1}$, compute: $\mathbf{P}^{k+1} = \max(0, \mathbf{P}^k - \alpha_k \nabla_{\mathbf{P}} \Delta(\mathbf{P}^k, \mathbf{W}^{k+1}, \mathbf{F}^{k+1}))$.
Normalize $\mathbf{P}$ by $\mathbf{P}_{ij}^{k+1} \leftarrow \frac{\mathbf{P}_{ij}^{k+1}}{\sum_j \mathbf{P}_{ij}^{k+1}}$.

until *convergence*;

Output: the optimized matrices $\mathbf{W}, \mathbf{P}$.

resents the distance between two distributions and minimizing it we impose the found partitions between pairs of related tasks to be consistent. The variables f_{ij}^{st} are flow variables as follows from the definition of EMD as a transportation problem [81].

In the proposed optimization problem there are no constraints on the $\mathbf{C}_t$ values. In this chapter we define the matrix $\mathbf{C} \in \boldsymbol{R}^{K \times d}$, $\mathbf{C} = [\mathbf{C}^{1\prime} \dots \mathbf{C}^{T\prime}]'$, $K = \sum_{t=1}^{T} K_t$, and we impose that the columns of $\mathbf{C}$ are a weighted sum of certain data points, *i.e.* $\mathbf{C} = \mathbf{PX}$ where $\mathbf{P} = \text{blkdiag}(\mathbf{P}^1, \dots, \mathbf{P}^T)$, $\mathbf{P} \in \boldsymbol{R}^{K \times N}$. In the following, for the sake of simplicity and easy interpretation, we consider only two tasks. The extension to T tasks is straightforward. Defining $\mathbf{F} = \text{diag}(f_{11} \dots f_{K_1 K_2})$, $\mathbf{F} \in \boldsymbol{R}^{K_1 K_2 \times K_1 K_2}$ and the block diagonal matrix $\mathbf{W} = \text{blkdiag}(\mathbf{W}^1, \mathbf{W}^2)$, $\mathbf{W} \in \boldsymbol{R}^{N \times K}$, we formulate the following optimization problem:

$$\Delta(\mathbf{P}, \mathbf{W}, \mathbf{F}) = \min_{\mathbf{P},\mathbf{W},\mathbf{F} \geq 0} \{\|\mathbf{X} - \mathbf{WPX}\|_F^2 + \lambda \text{tr}(\mathbf{MPXX}'\mathbf{P}'\mathbf{M}'\mathbf{F})\} \tag{4.2}$$

$$\begin{aligned} \text{s.t.} \quad & \|\mathbf{P}_{i.}^t\|_1 = 1, \ \ \forall i = 1, \dots, K \ \ \forall\, t = 1, 2 \\ & \text{tr}(\mathbf{I}_j \mathbf{F}) = \sum_{i=1}^{N} \mathbf{W}_{ij}, \ \ \forall j = 1, ..., K \\ & \text{tr}(\mathbf{F}) = 1 \end{aligned} \tag{4.3}$$

where: $\mathbf{I}_j \in \boldsymbol{R}^{K_1 K_2 \times K_1 K_2}$ and $\mathbf{M} \in \boldsymbol{R}^{K_1 K_2 \times K}$ are appropriately defined selection matrices. To solve the proposed optimization problem we develop an iterative optimization scheme described below. It is worth noting that our method can

Algorithm 5: Algorithm for solving (4.5).

Input: The data matrix $\mathbf{X}$, $\mathbf{E}, \mathbf{B}$, the parameter λ_2.

Set $\mathbf{Q} = \rho\mathbf{E}'\mathbf{E} + 2\mathbf{I} + 2\lambda_2\mathbf{B}$.
Compute Cholesky factorization of the matrix $\mathbf{Q}$.
for *j=1:d* **do**
  repeat
    Set $\mathbf{b}^k = \rho\mathbf{E}'\mathbf{q}^k - \mathbf{E}'\mathbf{p}^k + 2\mathbf{X}_{.j}$
    Update $\mathbf{\Pi}_{.j}$
    Solve $\mathbf{Q}[\mathbf{\Pi}_{.j}]^{k+1} = \mathbf{b}^k$
    Update $\mathbf{q}$ *using a soft thresholding operator*
    $\mathbf{q}^{k+1} = ST_{1/\rho}(\mathbf{E}[\mathbf{\Pi}_{.j}]^{k+1} + \frac{1}{\rho}\mathbf{p}^k)$
    Update $\mathbf{p}$
    $\mathbf{p}^{k+1} = \mathbf{p}^k + \rho(\mathbf{E}[\mathbf{\Pi}_{.j}]^{k+1} - \mathbf{q}^{k+1})$
  until *convergence*;

Output: The final centroid matrix $\mathbf{\Pi}$.

be kernelized, defining a feature mapping $\phi(\cdot)$ and the associated kernel matrix $\mathbf{K_X} = \phi(\mathbf{X})\phi(\mathbf{X})'$. The objective function of (4.2) becomes:

$$||\phi(\mathbf{X}) - \mathbf{WP}\ \phi(\mathbf{X})||_F^2 + \lambda \text{tr}(\mathbf{MP}\phi(\mathbf{X})\phi(\mathbf{X})'\mathbf{P}'\mathbf{M}'\mathbf{F}) =$$
$$\text{tr}(\mathbf{K_X} - \mathbf{2K_XP}'\mathbf{W}'+\mathbf{WPK_XP}'\mathbf{W}' + \lambda\mathbf{MPK_XP}'\mathbf{M}'\mathbf{F})$$

The update rules of the kernelized version of our method can be easily derived similarly to the linear case presented below using $\mathbf{K_X}$ instead of $\mathbf{X}'\mathbf{X}$.

4.3.2.1 Optimization

To solve (4.2), we first note that the optimal solution can be found by adopting an alternating optimization scheme, *i.e.* optimizing separately first with respect to $\mathbf{P}$ and then with respect to $\mathbf{W}$ and $\mathbf{F}$ jointly. In both cases, a non-negative least square problem with constraints arises, for which standard solvers can be employed. However, due to computational efficiency, in this chapter we consider an approximation of (4.2), replacing the constraints (4.3) with $\text{tr}(\mathbf{I}_j\mathbf{F}) = \mathbf{e}$, where $\mathbf{e} \in \boldsymbol{R}^{K_1K_2}$, $\mathbf{e}_i = \frac{1}{K_1}$, if $i \leq K_1$, $\mathbf{e}_i = \frac{1}{K_2}$ otherwise. This approximation implies that for each task the same number of datapoints is assigned to all the clusters. In this way a more efficient solver can be devised. Specifically, we

adopt an alternating optimization strategy, *i.e.* we optimize (4.2) separately with respect to $\mathbf{F}$, $\mathbf{W}$ and $\mathbf{P}$ until convergence, as explained in the following: **Step 1**: Fixed $\mathbf{W}, \mathbf{P}$, optimize $\mathbf{F}$ solving:

$$\begin{aligned} \min_{\mathbf{F}>\mathbf{0},\ \mathrm{tr}(\mathbf{F})=1} \quad & \mathrm{tr}(\mathbf{MPXX'P'M'F}) \\ \text{s.t.} \quad & \mathrm{tr}(\mathbf{I}_j\mathbf{F}) = \mathbf{e}, \quad \forall\, j = 1, ..., K_1 + K_2 \end{aligned} \tag{4.4}$$

This is a simple linear programming problem. It can be solved efficiently with standard solvers.

Step 2: Fixed $\mathbf{F}, \mathbf{P}$, optimize $\mathbf{W}$ solving:

$$\min_{\mathbf{W}>\mathbf{0}} ||\mathbf{X} - \mathbf{WPX}||_F^2$$

Following [60], we update $\mathbf{W}$ using a projected gradient method for bound-constrained optimization, *i.e.* $\mathbf{W}^{k+1} = \max(0, \mathbf{W}^k - \alpha_k \nabla_{\mathbf{W}} \Delta(\mathbf{P}^k, \mathbf{W}^k, \mathbf{F}^{k+1}))$, where $\nabla_{\mathbf{W}} \Delta(\mathbf{P}, \mathbf{W}, \mathbf{F}) = \mathbf{WPXX'P'} - \mathbf{XX'P'}$.

Step 3: Fixed $\mathbf{W}, \mathbf{F}$, optimize $\mathbf{P}$ solving:

$$\begin{aligned} \min_{\mathbf{P}>\mathbf{0}} \quad & ||\mathbf{X} - \mathbf{WPX}||_F^2 + \lambda \mathrm{tr}(\mathbf{MPXX'P'M'F}) \\ \text{s.t.} \quad & ||\mathbf{P}_{i.}^t||_1 = 1, \ \forall i \quad \forall\, t = 1, 2 \end{aligned}$$

Similarly to step 2, we update $\mathbf{P}$ using a projected gradient method for bound-constrained optimization, *i.e.* $\mathbf{P}^{k+1} = \max(0, \mathbf{P}^k - \alpha_k \nabla_{\mathbf{P}} \Delta(\mathbf{P}^k, \mathbf{W}^{k+1}, \mathbf{F}^{k+1}))$, where $\nabla_{\mathbf{P}} \Delta(\mathbf{P}, \mathbf{W}, \mathbf{F}) = \mathbf{W'WPXX'} - \mathbf{W'XX'} + \lambda \mathbf{M'FMPXX'}$. To account for constraints at each iteration we also normalize each row of $\mathbf{P}$, following the normalization invariance approach in [30].

The algorithm for solving (4.2) is summarized in Algorithm 4. Regarding the computational complexity, the cost of solving (4.2) with the iterative approach outlined in Algorithm 4 is dominated by the first step, *i.e.* by the linear programming problem in (4.4) which can be solved in polynomial time.

4.3.3 Convex Multi-task Clustering

Given the task specific training sets X^t, we propose to learn the sets of cluster centroids $\Pi^t = \{\boldsymbol{\pi}_1^t, \boldsymbol{\pi}_2^t, ..., \boldsymbol{\pi}_{N_t}^t\}, \boldsymbol{\pi}_i^t \in \boldsymbol{R}^d$, by solving the following optimization problem:

$$\min_{\boldsymbol{\pi}_i^t} \{\sum_{t=1}^{T}\sum_{i=1}^{N_t} \|\mathbf{x}_i^t - \boldsymbol{\pi}_i^t\|_2^2 + \lambda_t \sum_{t=1}^{T}\sum_{\substack{i,j=1 \\ j>i}}^{N_t} w_{ij}^t \|\boldsymbol{\pi}_i^t - \boldsymbol{\pi}_j^t\|_1 + \lambda_2 \boldsymbol{\Omega}_C(\Pi^t)\} \tag{4.5}$$

where:

$$\boldsymbol{\Omega}_C(\Pi^t) = \sum_{\substack{t,s=1 \\ s>t}}^{T} \gamma_{st} \sum_{i=1}^{N_t}\sum_{j=1}^{N_s} \|\boldsymbol{\pi}_i^t - \boldsymbol{\pi}_j^s\|_2^2$$

In (4.5) the first two terms guarantee that the data of each task are clustered: specifically with $\lambda_t = 0$ the found centroids are equal to the data-points while as λ_t increases the number of different centroids $\boldsymbol{\pi}_i^t$ reduces. The last term $\boldsymbol{\Omega}_c(\Pi^t)$ instead imposes the found centroids to be similar if the tasks are related. The relatedness between tasks is modelled by the parameter γ_{st} which can be set using an appropriate measure between distributions. We consider the Maximum Mean Discrepancy [11], defined as $\mathcal{D}(X^t, X^s) = \|\frac{1}{N_t}\sum_{i=1}^{N_t} \phi(\mathbf{x}_i^t) - \frac{1}{N_s}\sum_{i=1}^{N_s} \phi(\mathbf{x}_i^s)\|^2$ and we compute it using a linear kernel. We set $\gamma_{st} = e^{-\beta \mathcal{D}(X^t, X^s)}$ with β being a user-defined parameter ($\beta = 0.1$ in our experiments). The parameters w_{ij}^t are used to enforce datapoints in the same task to be assigned to the same cluster and can be set according to some *a-priori* knowledge or in a way such that the found partitions structure reflects the density of the original data distributions.

4.3.3.1 Optimization

To solve (4.5) we propose an algorithm based on the alternating direction method of multipliers [12]. We consider the matrix $\boldsymbol{\Pi} = [\boldsymbol{\Pi}^{1\prime}\ \boldsymbol{\Pi}^{2\prime}\ \ldots\ \boldsymbol{\Pi}^{T\prime}]', \boldsymbol{\Pi} \in \boldsymbol{R}^{N \times d}$, obtained concatenating the task-specific matrices $\boldsymbol{\Pi}^t = [\boldsymbol{\pi}_1^t\ \boldsymbol{\pi}_2^t\ ...\ \boldsymbol{\pi}_{N_t}^t]'$.

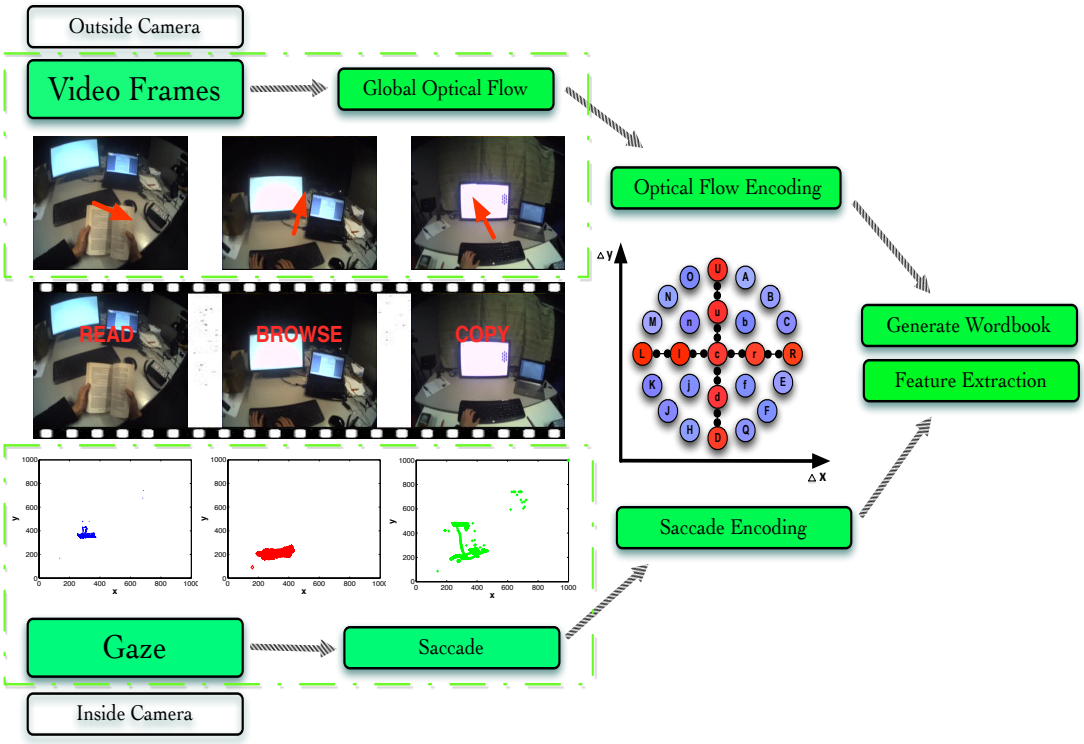


Figure 4.2: Feature extraction pipeline on the FPV office dataset. Some frames corresponding to the actions *read*, *browse* and *copy* are shown together with the corresponding optical flow features (top) and eye-gaze patterns depicted on the 2-D plane (bottom). It is interesting to observe the different gaze patterns among these activities.

The problem (4.5) can be solved considering d separate minimization subproblems (one for each column of $\mathbf{X}$) as follows:

$$\begin{aligned} \min_{\mathbf{q},\, \boldsymbol{\Pi}_{\cdot j}} \quad & \{\|\mathbf{X}_{\cdot j} - \boldsymbol{\Pi}_{\cdot j}\|_2^2 + \|\mathbf{q}\|_1 + \lambda_2 \|\mathbf{B}\boldsymbol{\Pi}_{\cdot j}\|_2^2\} \\ \text{s.t.} \quad & \mathbf{E}\boldsymbol{\Pi}_{\cdot j} - \mathbf{q} = 0 \end{aligned} \tag{4.6}$$

where $\mathbf{E}$ is a block diagonal matrix defined as $\mathbf{E} = \text{blkdiag}(\mathbf{E}^1, \mathbf{E}^2, \ldots, \mathbf{E}^T)$ and $\mathbf{E}^t \in \boldsymbol{R}^{|\mathcal{E}_t| \times N_t}$ is a matrix with $|\mathcal{E}_t| = \frac{N_t(N_t-1)}{2}$ rows. Each row is a vector of all zeros except in the position i where it has the value $\lambda_t w_{ij}^t$ and in the position j where it has the value $-\lambda_t w_{ij}^t$. Similarly the matrix $\mathbf{B} \in \boldsymbol{R}^{|\mathcal{B}| \times N}$, where $|\mathcal{B}| = \frac{T(T-1)}{2}$, imposes smoothness between the parameters of related tasks. A row of the matrix $\mathbf{B}$ is a vector with all zeros except in the terms corresponding to

datapoints of the t-th task which are set to γ_{st} and to the terms corresponding to datapoints of the s-th task which are all set to $-\gamma_{st}$. To solve (4.6) we consider the associated lagrangian:

$$L_\rho(\mathbf{\Pi}_{.j}, \mathbf{q}, \mathbf{p}) = \|\mathbf{X}_{.j} - \mathbf{\Pi}_{.j}\|_2^2 + \|\mathbf{q}\|_1 + \lambda_2 \|\mathbf{B}\mathbf{\Pi}_{.j}\|_2^2 + \mathbf{p}'(\mathbf{E}\mathbf{\Pi}_{.j} - \mathbf{q}) + \frac{\rho}{2} \left\|\mathbf{E}\mathbf{\Pi}_{.j} - \mathbf{q}\right\|_2^2$$

with $\mathbf{p}$ being the vector of augmented Lagrangian multipliers and ρ being the dual update step length. We devise an algorithm where three steps, corresponding to the update of the three variables $\mathbf{\Pi}_{.j}, \mathbf{q}, \mathbf{p}$, are performed.

Step 1: Update $\mathbf{\Pi}_{.j}$, given $\mathbf{q}, \mathbf{p}$ fixed, by solving:

$$\min_{\mathbf{\Pi}_{.j}} \; \|\mathbf{X}_{.j} - \mathbf{\Pi}_{.j}\|_2^2 + \|\mathbf{q}\|_1 + \lambda_2 \|\mathbf{B}\mathbf{\Pi}_{.j}\|_2^2 + \mathbf{p}'(\mathbf{E}\mathbf{\Pi}_{.j} - \mathbf{q}) + \frac{\rho}{2} \left\|\mathbf{E}\mathbf{\Pi}_{.j} - \mathbf{q}\right\|_2^2$$

Imposing the gradient with respect to $\mathbf{\Pi}_{.j}$ equal to 0, the update step is formulated as:

$$\mathbf{Q}[\mathbf{\Pi}_{.j}]^{k+1} = \mathbf{b}^k$$

where $\mathbf{Q} = \rho\mathbf{E}'\mathbf{E} + 2\mathbf{I} + 2\lambda_2\mathbf{B}$ and $\mathbf{b}^k = \rho\mathbf{E}'\mathbf{q}^k - \mathbf{E}'\mathbf{p}^k + 2\mathbf{X}_{.j}$. The computation of $\mathbf{\Pi}_{.j}$ involves solving a linear system. To solve it efficiently, we use Cholesky factorization and decompose $\mathbf{Q} = \mathbf{\Sigma}'\mathbf{\Sigma}$. In practice, at each iteration, we solve two linear systems: $\mathbf{\Sigma}'\mathbf{g} = \mathbf{b}^k$ and $\mathbf{\Sigma}\mathbf{\Pi}_{.j} = \mathbf{g}$. Since $\mathbf{\Sigma}$ is an upper triangular matrix, solving them is typically very efficient. **Step 2:** Update $\mathbf{q}$, given $\mathbf{\Pi}_{.j}, \mathbf{p}$ fixed, by solving:

$$\min_{\mathbf{q}} \quad \|\mathbf{q}\|_1 - \mathbf{p}'\mathbf{q} + \frac{\rho}{2} \left\|\mathbf{E}\mathbf{\Pi}_{.j} - \mathbf{q}\right\|_2^2$$

Neglecting the constant terms, the update step is:

$$\mathbf{q}^{k+1} = \arg\min_{\mathbf{q}} \frac{1}{2} \left\| \mathbf{q} - \mathbf{E}[\mathbf{\Pi}_{.j}]^{k+1} - \frac{1}{\rho}\mathbf{p}^k \right\|_2^2 + \frac{1}{\rho} \|\mathbf{q}\|_1$$

This equation has a closed-form solution. Defining the soft thresholding operator $ST_\lambda(x) = \text{sign}(x)\max(|x| - \lambda, 0)$ the update step becomes:

$$\mathbf{q}^{k+1} = ST_{1/\rho}(\mathbf{E}[\mathbf{\Pi}_{\cdot j}]^{k+1} + \frac{1}{\rho}\mathbf{p}^k)$$

Step 3: Update $\mathbf{p}$, given $\mathbf{\Pi}_{\cdot j}$, $\mathbf{q}$ fixed, with the equation:

$$\mathbf{p}^{k+1} = \mathbf{p}^k + \rho(\mathbf{E}[\mathbf{\Pi}_{\cdot j}]^{k+1} - \mathbf{q}^{k+1})$$

We summarize our approach in Algorithm 5. Regarding the computational complexity of Algorithm 5, the most computationally expensive step is the Cholesky matrix factorization ($O(N^3)$). However, the Cholesky factorization is performed only once. In the inner loop, for each dimension $j = 1, \ldots, d$, each iteration involves solving one linear system ($O(N^2)$) and a soft-thresholding operation ($O(\sum_t |\mathcal{E}^t|)$).

4.3.4 Features Extraction in Egocentric Videos

The growing interest in the vision community towards novel approaches for FPV analysis has motivated the creation of several publicly available datasets (see [92] for a recent survey). In this chapter we consider two of them, the FPV office dataset [74] and the FPV home dataset [77].

Due to the large variability of visual data collected from wearable cameras there exist no standard feature descriptors. While in some situations extracting simple motion information, *e.g.* by computing the optical flow, may suffice [74], in other cases motion patterns may be too noisy and other kind of information (*e.g.* presence/absence of objects) must be exploited. In this chapter we demonstrate that, independently from the employed feature descriptors, MTC is an effective strategy for recognizing everyday activities. We now describe the adopted feature representations respectively for the considered office and home scenarios.

4.3.4.1 FPV office dataset

The FPV office dataset [74] consists of five common activities in an office environment (*reading a book, watching a video, copying text from screen to screen, writing sentences on paper* and *browsing the internet*). Each action was performed by five subjects, who were instructed to execute each task for about two minutes, while 30 seconds intervals of void class were placed between target tasks. To provide a natural experimental setting, the void class contains a wide variety of actions such as conversing, singing and random head motions. The sequence of five actions was repeated twice to induce interclass variance. The dataset consists of over two hours of data, where the video from each subject is a continuous 25-30 minutes video.

We follow [74] and extract features describing both the eye motion (obtained by the inside-looking camera) and the head and body motion (computed processing the outside camera's stream). To calculate the eye motion features, we consider the gaze coordinates provided in the dataset and smooth them applying a median filter. Then the continuous wavelet transform is adopted for saccade detection separately on the x and y motion components [13]. The resulting signals are quantized according to magnitude and direction and are coded with a sequence of discrete symbols. To analyze the streams of the output camera, for each frame the global optical flow is computed by tracking corner points over consecutive frames and taking the mean flow in the x and y directions. Then, the optical flow vectors are quantized according to magnitude and direction with the same procedure adopted in the eye motion case. The obtained sequences of symbols are then processed to get the final video clip descriptors. We use a temporal sliding window approach to build an n-gram dictionary over all the dataset. Then each video is divided into segments corresponding to 15 seconds, each of them representing a video clip. For each sequence of symbols associated to a video clip, a histogram over the dictionary is computed. The final feature descriptor $\mathbf{x}_i$ is calculated by considering some statistics over the clip histogram

and specifically computing the maximum, the average, the variance, the number of unique n-grams, and the difference between maximum and minimum count. Fig.4.2 shows the feature extraction pipeline.

4.3.4.2 FPV home dataset

The FPV home dataset [77] contains videos recorded from chest-mounted cameras by 20 different users. The users perform 18 non-scripted daily activities in the house, like *brushing teeth, washing dishes,* or *making tea*. The length of the videos is in the range of 20-60 minutes. The annotations about the presence of 42 relevant objects (*e.g.* kettle, mugs, fridge) and about temporal segmentation are also provided.

In this chapter we adopt the same object-centric approach proposed in [77], *i.e.* to compute features for each video clip we consider the output of several object detectors. We use the pre-segmented video clips and the active object models in [77]. Active object models are introduced to exploit the fact that objects may look different when being interacted with (*e.g.* open and close fridge). Therefore in [77] additional detectors are trained using a subset of training images depicting the object appearance when objects are used by people. To obtain object-centric features for each frame a score for each object model and each location is computed. The maximum scores of all the object models are used as frame features. To compute the final clip descriptor $\mathbf{x}_i$, two approaches are adopted: one based on "bag of features" (accumulating frame features over time) and the other based on temporal pyramids. The temporal pyramid features are obtained concatenating multiple histograms constructed with accumulation: the first is a histogram over the full temporal extent of a video clip, the next is the concatenation of two histograms obtained by temporally segmenting the video into two parts, etc.

4.4 Experimental Results

In this section, we first conduct experiments on synthetic data to demonstrate the advantages of the proposed MTC approach over traditional single task learning methods. Then, we apply our MTC algorithms to FPV data showing their effectiveness for recognizing everyday activities.

In the experiments, we compare our methods, *i.e.* EMD Multi-task Clustering with linear and gaussian kernel and Convex Multi-task Clustering (here denoted as EMD-MTC, KEMD-MTC and CMTC, respectively), with single task clustering approaches. Specifically we consider k-means (KM), kernel k-means (KKM), convex (CNMF) and semi-nonnegative matrix factorization (SemiNMF) [27]. We also consider recent multi-task clustering algorithms such as the SemiEMD-MTC proposed in [129], its kernel version KSemiEMD-MTC and the LS-MTC method in [40]. For all the methods (with the exception of CMTC which relies on convex optimization) ten runs are performed, corresponding to different initializations conditions. Averaging over multiple iterations is typical when considering non-convex optimization problems for clustering, such as in case of the popular k-means. The average results are shown. In CMTC the parameters λ_t are varied in order to obtain the desired number of clusters. The value of the regularization parameters of our approaches (λ for the methods based on EMD regularization and λ_2 for CMTC) are set in the range $\{0.01, 0.1, 1, 10, 100\}$. As evaluation metrics, we adopt the clustering accuracy (ACC) and the normalized mutual information (NMI), as they are widely used in the literature.

4.4.1 Synthetic data experiments

In the synthetic data experiments we consider $T = 4$ different tasks. Each task contains 4 clusters as shown in Fig.4.3. The input data $\mathbf{x}_i^t \in R^d$ ($d = 2$) for the four clusters are generated from multivariate normal distributions $\mathcal{N}(\mu, \sigma)$,

Table 4.1: Parameters used in the synthetic data experiments.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
	μ_1	μ_2	μ_3	μ_4
Task 1	(0, 0)	(1, 1)	(-1, 1)	(1, -1)
Task 2	(-0.2, -0.22)	(1, 1.04)	(-1, 0.95)	(1.2, -0.83)
Task 3	(0.02, 0)	(1.04, 1)	(-0.95, 1)	(1.03, -1)
Task 4	(-0.22, -0.22)	(1.04, 1.04)	(-0.95, 0.95)	(1.23, -0.83)
σ	(0.1, 0.1)	(0.2, 0.4)	(0.1, 0.2)	(0.4, 0.2)

Table 4.2: FPV office dataset: comparison of different methods using saccade (S), motion (M) and S+M features.

	ACC			NMI		
	S	M	S+M	S	M	S+M
KM	0.230	0.216	0.257	0.029	0.021	0.045
SemiNMF [27]	0.320	0.303	0.358	0.149	0.131	0.166
SemiEMD-MTC [129]	0.371	0.349	0.415	0.229	0.209	0.259
LSMTC [40]	0.286	0.261	0.335	0.043	0.031	0.071
CNMF [27]	0.328	0.301	0.357	0.152	0.139	0.170
EMD-MTC	0.389	0.363	0.442	0.239	0.221	0.273
CMTC ($\lambda_2 = 0$)	0.367	0.346	0.413	0.224	0.209	0.244
CMTC	0.425	0.401	0.468	0.259	0.238	0.305
KKM	0.345	0.316	0.377	0.159	0.152	0.185
KSemiEMD-MTC [129]	0.387	0.359	0.432	0.241	0.228	0.287
KEMD-MTC	0.436	0.419	0.485	0.268	0.244	0.311

as shown in Table 4.1, in order to obtain correlated clusters for the different tasks. For each task and each cluster 10 samples are generated for training and 20 are used to set the regularization parameters. For CMTC we set the weights $w_{ij}^t = e^{-\|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2}$ if $e^{-\|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2} \leq \theta$ and $w_{ij}^t = 0$ otherwise. This aims to enforce that the discovered partitions reflect the density of the original data distributions.

We compared the proposed methods with other state-of-the-art approaches. Fig.4.4 reports the average accuracy and NMI. The higher numbers indicate better performance. From Fig.4.4 it is evident that our multi-task approaches

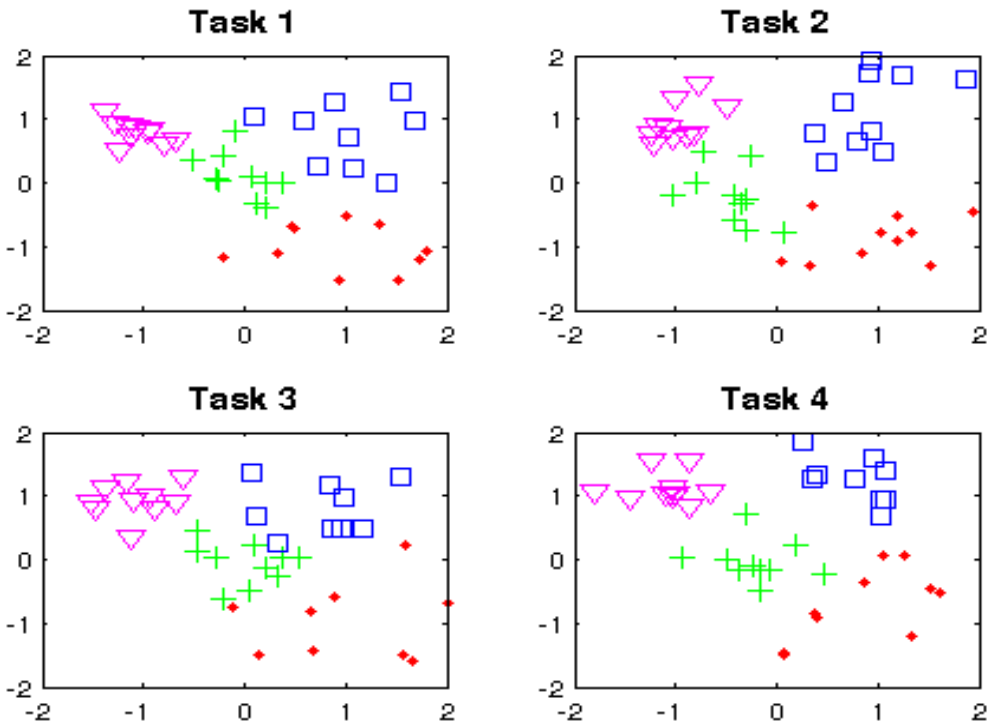


Figure 4.3: Samples generated in the synthetic data experiments (different colors represent different clusters).

significantly outperform the single-task methods, both when a linear kernel is used (*e.g.* EMD-MTL and CMTC achieve higher accuracy than KM), and in the nonlinear case (KEMD-MTC outperforms KKM). The proposed algorithms also achieve higher accuracy than recent multi-task clustering methods, *i.e.* KSemiEMD-MTC [129] and LS-MTC [40].

4.4.2 FPV Results

In this subsection, we present the experimental results on the FPV office dataset and the FPV home dataset, respectively.

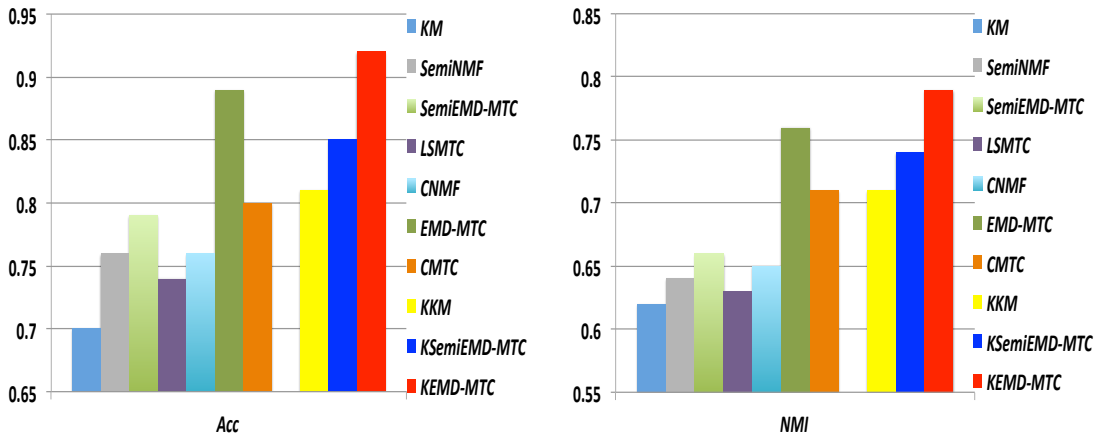


Figure 4.4: Clustering results on synthetic data for different methods. Methods based on linear kernel are separated from those with Gaussian kernel. (Figure is best viewed in color).

4.4.2.1 FPV office dataset

We consider $T = 5$ tasks, as the FPV office dataset [74] contains videos corresponding to five people. As each datapoint corresponds to a video clip in this dataset, we set the parameters w_{ij}^t in CMTC in order to enforce temporal consistency, *i.e.* for each task t, $w_{ij}^t = 1$ if the features vectors $\mathbf{x}_i^t$ and $\mathbf{x}_j^t$ correspond to temporal adjacent video clips, otherwise $w_{ij}^t = 0$.

Table 4.2 compare different clustering methods when different types of features are employed, *i.e.* only saccade, only motion and saccade+motion features. The last three rows correspond to methods which employ a non-linear kernel. From Table 4.2, several observations can be made. First, independently on the adopted features representation, multi-task clustering approaches always perform better than single task clustering methods (*e.g.* SemiEMD-MTC outperforms SemiNMF, EMD-MTC provides higher accuracy than CNMF, a value of λ_2 greater than 0 leads to an improvement in accuracy and NMI in CMTC). Confirming the findings reported in [74], we also observe that combining motion and saccade features is advantageous with respect to considering each single feature representation separately. Noticeably, our methods are among the best performers, with KEMD-MTC reaching the highest values of accuracy and

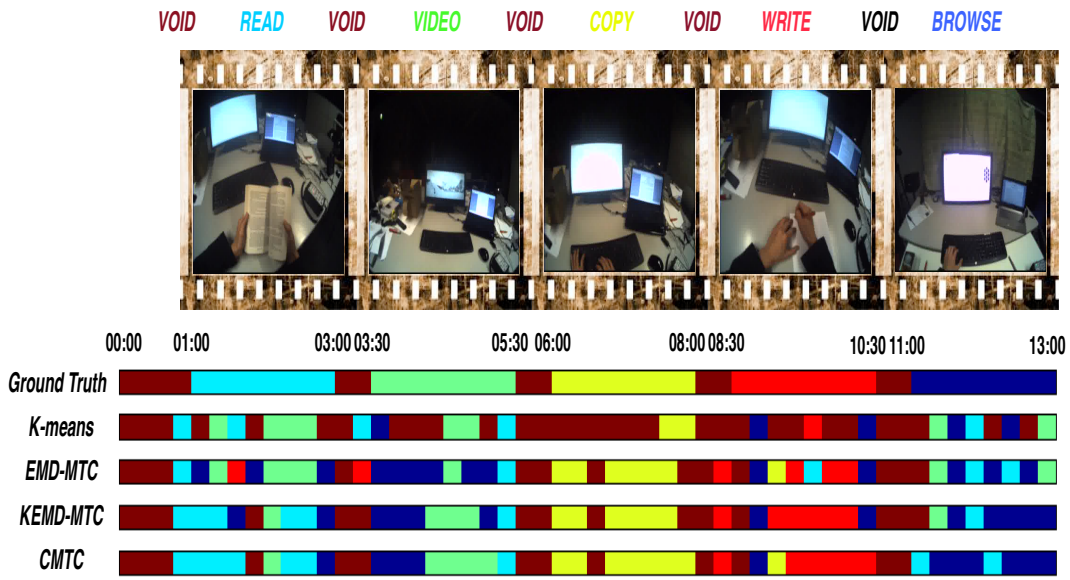


Figure 4.5: FPV Office dataset. Temporal video segmentation on the second sequence of subject-3 (13 minutes): comparison of different methods. (Best viewed in color).

NMI. This is somehow expected probably due to both the use of kernels and the adoption of the multi-task learning paradigm. Moreover, CMTC outperforms EMD-MTC by up to 4% which means that incorporating information about temporal consistency in the learning process is beneficial. Furthermore, in this case the use of Maximum Mean Discrepancy may capture better the relationship among tasks with respect to EMD. Fig.4.5 shows some qualitative temporal segmentation results on the second sequence of subject-3. In this case for example the CMTC method outperforms all the other approaches and the importance of enforcing temporal consistency among clips is evident.

Finally, Fig.4.6 shows the confusion matrices associated to our methods KEMD-MTC and CMTC. Examining the matrix associated to KEMD-MTC, we observe that the *void, copy* and *write* actions achieve relative high recognition accuracies compared with the *video* and *browse* actions. It is also interesting to note that 25% and 17% of the *video* actions are recognized as *browse* actions for KEMD-MTC and CMTC respectively, because of the similarity among motion and eye-gaze patterns.

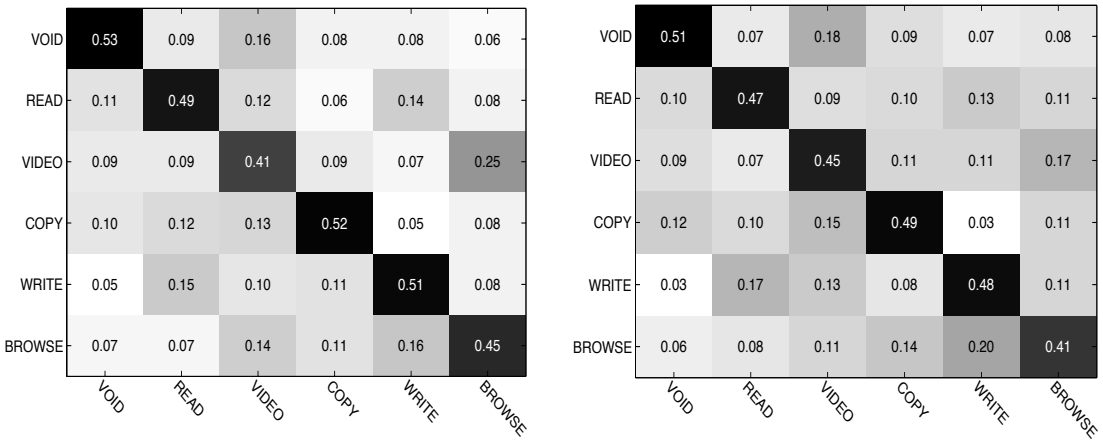


Figure 4.6: FPV Office dataset. Confusion matrices using saccade+motion features obtained with (left) KEMD-MTC and (right) CMTC methods.

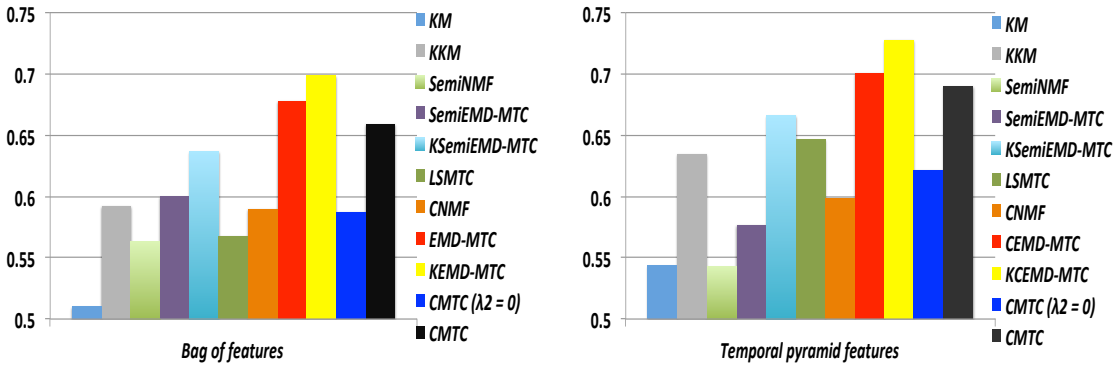


Figure 4.7: Comparison of different methods using (left) bag of features and (right) temporal pyramid features on FPV home dataset. (Figure is best viewed in color).

4.4.2.2 FPV home dataset

In the FPV home dataset [77] there are 18 different non-scripted activities. Since each person typically performs a small subset of the 18 activities, in our experiments we consider a series of three tasks problems, selecting videos associated to three randomly chosen users but imposing the condition that videos corresponding to the three users should have at least three activities in common. We perform 10 different runs. In this series of experiments, we did not cluster video clips of fixed size as in the office dataset, but we consider the pre-segmented clips as provided with the dataset. In this scenario, it does not make sense to

set w_{ij}^t as in CMTC to model temporal consistency. Therefore, as for in the synthetic data experiments, we set $w_{ij}^t = e^{-\|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2}$ if $e^{-\|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2} \leq \theta$ and $w_{ij}^t = 0$ otherwise.

Fig.4.7 shows the results (average accuracy) obtained with different clustering methods for both the bag-of-words and the temporal pyramid features representation. From Fig.4.7 it is evident that the MTC approaches outperforms their single task version (*e.g.* CMTC outperforms CMTC with $\lambda_2 = 0$, EMD-MTC outperforms CNMF, SemiEMD-MTC outperforms SemiNMF). On the other hand, our algorithms based on EMD regularization and CMTC achieve a considerably higher accuracy with respect to all the other methods. Fig.4.8 shows some temporal segmentation results on a sequence of the FPV home dataset comparing KM with the proposed methods. As discussed above, pre-segmented clips of different duration are considered here.
Finally, we investigate the effect of different values of the regularization parameters λ in (4.2) for EMD-MTC, λ_t and λ_2 in (4.5) for CMTC on clustering performance. As shown in Fig.4.9, independently from the adopted feature representation, the accuracy values are sensitive to varying λ. Fig.4.9 shows that choosing a value of $\lambda = 0.1$ in EMD-MTC and KEMD-MTC always leads to similar or superior performance with respect to adopting a single-task clustering approach ($\lambda = 0$). The value $\lambda = 0.1$ corresponds to the results reported in Fig.4.7. This clearly confirms the advantage of using a MTC approach for FPV analysis. Similar observations can be drawn in the case of CMTC. In Fig.4.10 we analyze how the accuracy changes at varying λ_t and λ_2. Note that in our previous experiments the parameters λ_t are fixed independently for each task according to the desired number of clusters. In this experiment instead we show that, independently from the chosen values for λ_t (*i.e.* the number of clusters) the best performance is typically obtained for $\lambda_2 \geq 0.1$, *i.e.* when the coherence between partitions of different tasks is enforced. For example, for temporal pyramid features, the higher accuracy is usually given by $\lambda_2 = 1$.

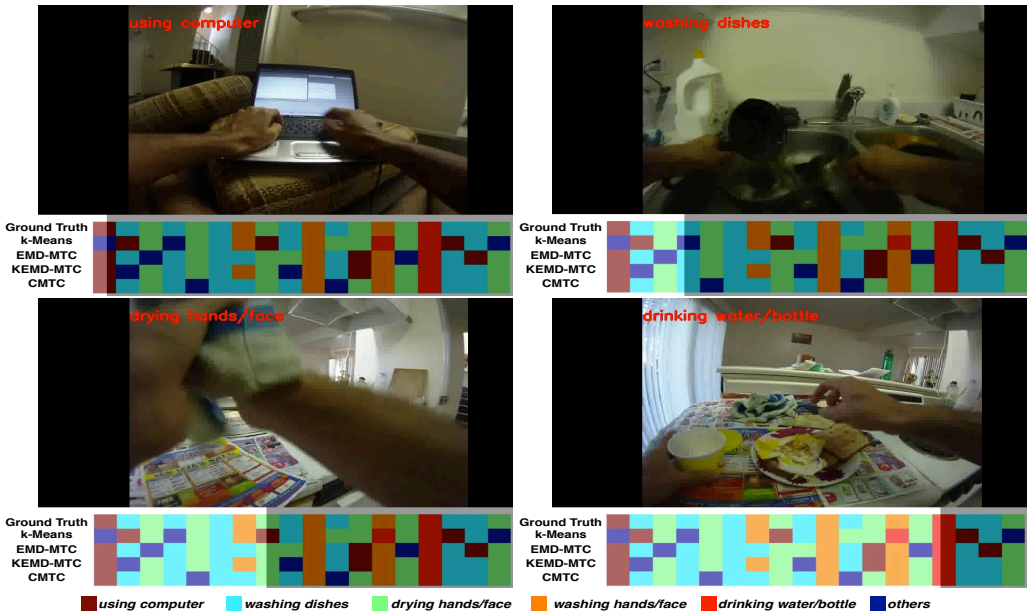


Figure 4.8: Temporal video segmentation on a sequence of the FPV home dataset. (The edge of the shaded area at the bottom of each subfigure indicates the current frame).

4.4.3 Discussion

In this chapter we address the problem of automatically discovering activities of daily living from first-person videos. Currently, few datasets are publicly available for this task and, according to the recent survey in [92], the two datasets we consider [74, 77] are the only ones suitable. The other datasets focus on different applications, *e.g.* food preparation or analysis of social interactions, and often do not have videos recorded from multiple users, as required by the proposed framework.

Regarding previous works using the same datasets [74, 77], it is worth noting that we consider an unsupervised setting. Previous works focused on a supervised scenario and therefore use different evaluation metrics. While a direct comparison is not possible, it is reasonable to expect that their methods are more accurate than our approach since they use labelled data for learning. However, recognizing everyday activities in absence of annotated data is especially important to automatically analyze videos recorded from wearable cameras.

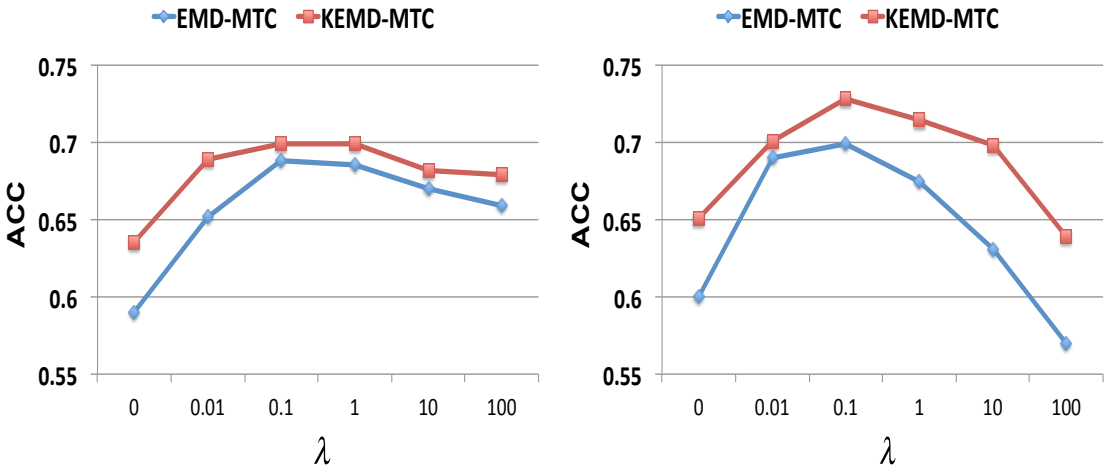


Figure 4.9: FPV home dataset: performance variations of EMD-MTC and KEMD-MTC at different values of λ using (left) bag of features and (right) temporal pyramid features.

As stated in the introduction, the proposed multi-task clustering approach is general and can be used in other applications. For example, our framework naturally applies to the problem of activity of daily living analysis when traditional cameras are used as an alternative to wearable sensors [56, 71, 80, 102].

4.5 Conclusions

In this chapter, we proposed a multi-task clustering framework to tackle the challenging problem of egocentric activity recognition. Oppositely to many previous works, we focused on the unsupervised setting and we presented two novel MTC algorithms: Earth Movers Distance Multi-Task Clustering and Convex Multi-task Clustering. We extensively evaluated the proposed methods on synthetic data and on two real world FPV datasets, clearly demonstrating the advantages of sharing informations among related tasks over traditional single task learning algorithms. Comparing the proposed methods, KEMD-MTC achieves the best performance, while CMTC is particularly advantageous when some *a-priori* knowledge about the data relationship is available. For example, in this chapter we consider embedding temporal information about video clips

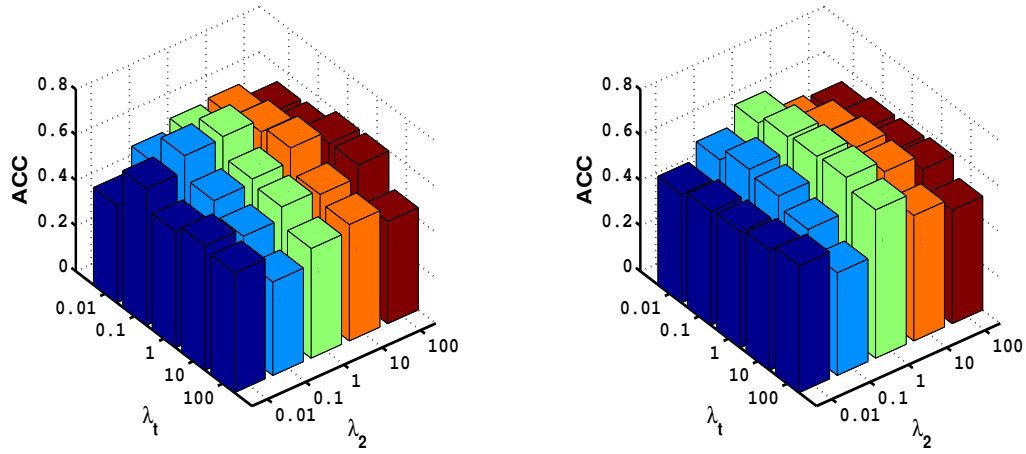


Figure 4.10: Sensitivity study of parameters λ_t and λ_2 in CMTC using (left) bag of features and (right) temporal pyramid features.

but the CMTC method also permits to integrate other information about task dependencies by defining an appropriate matrix $\mathbf{B}$ (*e.g.* people performing the same activities in the same rooms may correspond to closely related tasks with respect to people operating in different rooms). Future work will focus on improving our MTC algorithms (*e.g.* by detecting outlier tasks) and on testing the effectiveness of the proposed methods for other vision applications.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we address several image and video recognition problems under the framework of learning with shared information.

- In Chapter 2, we propose an active transfer learning framework which explicitly accounts for ambiguous labels provided by the domain expert. Moreover, we also extend traditional active learning from binary classification to a multi-class setting through error-correcting output coding.

- In Chapter 3, we investigate how to automatically infer painting styles from the perspective of dictionary learning. In particular, we propose a novel multi-task dictionary learning approach to address this problem. We also evaluate our approach on other image recognition datasets.

- In Chapter 4, we propose a unsupervised multi-task clustering framework to tackle the challenging problem of egocentric activity recognition. We focus on the unsupervised setting and present two novel multi-task clustering algorithms, Earth Movers Distance Multi-Task Clustering and Convex Multi-task Clustering.

To summarize, the contributions of this thesis are as follows:

- We explore several challenge recognition problems, such as human daily activity recognition under the framework of learning with shared information.
- We develop several novel machine learning algorithms under the framework of learning with shared information, such as active transfer learning, multi-task clustering, multi-task dictionary learning.
- Our algorithms outperform other state-of-the-art algorithms in several image and video recognition problems.
- All the proposed algorithms are general framework, potentially applicable to other computer vision and pattern recognition problems.

5.2 Future Work

In the future, we will continue our research with the following possible directions:

- Our models can be extended to other applications, such as human action recognition, video segmentation, crowd analysis.
- Deep learning based framework can be explored to address these applications. We can build separate networks for each domain, or share the same network for different domains. We can also explore which specific layers can be shared, etc.

Bibliography

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE TIP*, 54(11):4311–4322, 2006.

[2] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *JMLR*, 1(1):113–141, 2000.

[3] Oisin Mac Aodha, Neill D.F. Campbell, Jan Kautz, and Gabriel J. Brostow. Hierarchical subquery evaluation for active learning on a graph. In *CVPR*, 2014.

[4] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *NIPS*, 2006.

[5] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In *NIPS*, 2007.

[6] A. Argyrious and T. Evegenious. Multi-task feature learning. In *NIPS*, 2007.

[7] Muhammad Amir As'ari and Usman Ullah Sheikh. Vision based assistive technology for people with dementia performing activities of daily living (adls): an overview. In *Int. Conf. on Digital Image Processing*, 2012.

[8] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[9] Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. *Journal of Machine Learning Research*, 20:97–112, 2011.

[10] E. Bonilla, K.M. Chai, and C. Williams. Multi-task gaussian process prediction. In *NIPS*, 2008.

[11] K. Borgwardt, A. Gretton, M. Rasch, H.-P. Kriegel, B. Schoelkopf, and A. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):1–9, 2006.

[12] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

[13] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):741–753, 2011.

[14] Gustavo Carneiro. Artistic image analysis using graph-based learning approaches. *IEEE TIP*, 22(8):3168–3178, 2013.

[15] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[16] Pierluigi Casale, Oriol Pujol, and Petia Radeva. Human activity recognition from accelerometer data using a wearable device. In *Pattern Recognition and Image Analysis*, pages 289–296. Springer, 2011.

[17] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. 2001.

[18] Xiaojun Chang, Feiping Nie, Yi Yang, and Heng Huang. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, 2014.

[19] Xiaojun Chang, Feiping Nie, Yi Yang, and Heng Huang. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, 2014.

[20] Xiaojun Chang, Fieping Nie, Zhigang Ma, Yi Yang, and Xiaofang Zhou. A convex formulation for spectral shrunk clustering. In *AAAI*, 2015.

[21] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *SIGKDD*, 2011.

[22] Florin Cutzu, Riad Hammoud, and Alex Leykin. Distinguishing paintings from photographs. *CVIU*, 100:294–273, 2005.

[23] Wenyuan Dai, Qiang Yang, Guirong Xue, and Yong Yu. Boosting for transfer learning. In *ICML*, 2007.

[24] W.Y. Dai, Q. Yang, and Y. Yu. Boosting for transfer learning. In *ICML*, 2007.

[25] M. Daldoss, N. Piotto, N. Conci, and F.G.B. De Natale. Learning and matching human activities using regular expressions. In *ICIP*, 2010.

[26] H. Daume. Frustratingly easy domain adaptation. In *ACL*, 2007.

[27] Chris Ding, Tao Li, and Michael I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, 2010.

[28] Jun Du and Charles X. Ling. Active learning with human-like noisy oracle. In *ICDM*, 2010.

[29] L.X. Duan, D. Xu, and I. W. Tsang. Visual event recognition in videos by learning from web data. In *CVPR*, 2010.

[30] J. Eggert and E. Korner. Sparse coding and NMF. *Neural Networks*, 4:2529–2533, 2004.

[31] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representation over learned dictionaries. *IEEE TIP*, 15(12):3736–3745, 2006.

[32] Ehsan Elhamifar, Guillermo Sapiro, and Shankar Sastry. A convex optimization framework for active learning. In *ICCV*, 2013.

[33] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi-task learning. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.

[34] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *IEEE International Conference on Computer Vision*, 2011.

[35] Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, 2012.

[36] Alireza Fathi and James M. Rehg. Social interactions: A first-person perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[37] Y. Freund and R. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[38] Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. In *NIPS*, 2010.

[39] A. Gretton, K. Borgwardt, and B. Scholkopt. A kernel method for the two-sample-problem. In *NIPS*, 2006.

[40] Quanquan Gu and Jie Zhou. Learning the shared subspace for multi-task clustering and transductive transfer classification. In *IEEE International Conference on Data Mining*, 2009.

[41] Yahong Han, Fei Wu, Dacheng Tao, Jian Shao, Yueting Zhuang, and Jianmin Jiang. Sparse unsupervised dimensionality reduction for multiple view data. *TCSVT*, 22:1485–1496, 2012.

[42] Yahong Han, Fei Wu, Yueting Zhuang, and Xiaofei He. Multi-label transfer learning with sparse representation. *TCSVT*, 20:1110–1121, 2010.

[43] Yahong Han, Yi Yang, Zhigang Ma, Haoquan Shen, Nicu Sebe, and Xiaofang Zhou. Image attribute adaptation. *TMM*, 16:1115–1126, 2014.

[44] S.C.H. Hoi, R. Jin, and M.R. Lyu. Large-scale text categorization by batch mode active learning. In *WWW*, 2006.

[45] Gang Hua, Chengjiang Long, Ming Yang, and Yan Gao. Collaborative active learning of a kernel machine ensemble for recognition. In *ICCV*, 2013.

[46] J. Huang, A. Smola, and B. Scholkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007.

[47] James M Hughes, Daniel J Graham, and Daniel N Rockmore. Quantification of artistic style through sparse coding analysis in the drawings of pieter bruegel the elder. *Proceedings of the National Academy of Sciences*, 107(4):1279–1283, 2010.

[48] L. Jacob, F. Bach, and J. Vert. Clustered multi-task learning: A convex formulation. In *Conference on Advances in Neural Information Processing Systems*, 2008.

[49] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *Conference on Advances in Neural Information Processing Systems*, 2010.

[50] T. Kanade and M. Hebert. First-person vision. *Proceedings of the IEEE*, 100(8):2442–2453, 2012.

[51] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *BMVC*, 2014.

[52] Shu Kong and Donghui Wang. A dictionary learning approach for classification: separating the particularity and the commonality. In *ECCV*, 2012.

[53] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011.

[54] Brian Kulis and Michael I Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *International Conference on Machine Learning*, 2012.

[55] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In *NIPS*, 2006.

[56] Jinna Lei, Xiaofeng Ren, and Dieter Fox. Fine-grained kitchen activity recognition using RGB-D. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2012.

[57] Jia Li, Lei Yao, Ella Hendriks, and James Z. Wang. Rhythmic brushstrokes distinguish van Gogh from his contemporaries: Findings via automated brushstroke extraction. *IEEE TPAMI*, 34(6):1159–1176, 2012.

[58] Xin Li and Yuhong Guo. Multi-level adaptive active learning for scene classification. In *ECCV*, 2014.

[59] Lucy Liang and Kristen Grauman. Beyond comparing image pairs: Setwise active learning for relative attributes. In *CVPR*, 2014.

[60] C-J Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19:2756–2779, 2007.

[61] G. Liu, Y. Yan, E. Ricci, Y. Yang, Y. Han, S. Winkler, and N. Sebe. Inferring painting style with multi-task dictionary learning. In *International Joint Conferences on Artificial Intelligence*, 2015.

[62] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[63] Yong Luo, Dacheng Tao, Bo Geng, Chao Xu, and S.J. Maybank. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE Transactions on Image Processing*, 22(2):523–536, 2013.

[64] Zhigang Ma, FeiPing Nie, Yi Yang, Jasper Uijlings, and Nicu Sebe. Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Transactions on Multimedia*, 14(4):1021–1030, 2012.

[65] Behrooz Mahasseni and Sinisa Todorovic. Latent multitask learning for view-invariant action recognition. In *IEEE International Conference on Computer Vision*, 2013.

[66] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *ICML*, 2009.

[67] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Discriminative learned dictionaries for local image analysis. In *CVPR*, 2008.

[68] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Discriminative learned dictionaries for local image analysis. In *CVPR*, 2008.

[69] Oded Maron and Tomas Lozano-Perez. A framework for multiple-instance learning. In *NIPS*, 1998.

[70] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera Paredes. Sparse coding for multitask and transfer learning. In *ICML*, 2013.

[71] Ross Messing, Chris Pal, and Henry Kautz. Activity recognition using the velocity histories of tracked keypoints. In *IEEE International Conference on Computer Vision*, 2009.

[72] L. Mihalkova, T. Huynh, and R.J. Mooney. Mapping and revising markov logic networks for transfer learning. In *AAAI*, 2007.

[73] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, 2013.

[74] Keisuke Ogaki, Kris M Kitani, Yusuke Sugano, and Yoichi Sato. Coupling eye-motion and ego-motion features for first-person activity recognition. In *CVPR Workshop on Egocentric Vision*, 2012.

[75] Aghazadeh Omid, Sullivan Josephine, and Carlsso Stefan. Novelty detection from an egocentric perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[76] S. Jialin Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, Vol 22, NO 10*, 22(10):1345–1359, 2010.

[77] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[78] Yair Poleg, Chetan Arora, and Shmuel Peleg. Temporal segmentation of egocentric videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[79] Anoop Kolar Rajagopal, Ramanathan Subramanian, Elisa Ricci, Radu L Vieriu, Oswald Lanz, Nicu Sebe, et al. Exploring transfer learning approaches for head pose classification from multi-view surveillance images. *IJCV*, 109(1-2):146–167, 2014.

[80] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[81] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *IEEE International Conference on Computer Vision*, 1998.

[82] Paul Ruvolo and Eric Eaton. Online multi-task learning via sparse dictionary optimization. In *AAAI*, 2014.

[83] M. S. Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[84] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[85] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[86] Lior Shamir, Tomasz Macura, Nikita Orlov, D Mark Eckley, and Ilya G Goldberg. Impressionism, expressionism, surrealism: Automated recog-

nition of painters and schools of art. *ACM Transactions on Applied Perception (TAP)*, 7(2):8, 2010.

[87] V.S. Sheng, F. Provost, and P.G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD*, 2008.

[88] X.X. Shi, W. Fan, and J.T. Ren. Actively transfer domain knowledge. In *ECML*, 2008.

[89] Yasuhiro Sogawa, Tsuyoshi Ueno, Yoshinobu Kawahara, and Takashi Washio. Active learning for noisy oracle via density power divergence. *Neural Networks*, 46:133–143, 2013.

[90] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *SIGMOD*, 2013.

[91] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Richang Hong. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *ACM Multimedia*, 2011.

[92] Sibo Song, Vijay Chandrasekhar, Ngai-Man Cheung, Sanath Narayan, Liyuan Li, and Joo-Hwee Lim. Activity recognition in egocentric lifelogging videos. In *Int. Workshop on Mobile and Egocentric Vision, Asian Conference on Computer Vision*, 2014.

[93] Rainer Stiefelhagen, Rachel Bowers, and G. Fiscus Jonathan. Multimodal technologies for perception of humans, CLEAR. 2007.

[94] Emmanuel Munguia Tapia, Stephen S. Intille, and Kent Larson. Activity recognition in the home using simple and ubiquitous sensors. In *Pervasive Computing*, pages 158–175, 2004.

[95] Ekaterina Taralova, Fernando De la Torre, and Martial Hebert. Source constrained clustering. In *IEEE International Conference on Computer Vision*, 2011.

[96] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *ICML*, 2000.

[97] Antonio Torralba and Alexei Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[98] Pavan Turaga, Rama Chellappa, Venkatramana S Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.

[99] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[100] Xuezhi Wang, Tzu-Kuo Huang, and Jeff Schneider. Active transfer learning under model shift. In *ICML*, 2014.

[101] Ying Wang and Masahiro Takatsuka. SOM based artistic styles visualization. In *ICME*, 2013.

[102] Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, et al. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127:14–30, 2014.

[103] R. Yan, J. Yang, and A. G. Hauptmann. Automatically labeling video data using multi-class active learning. In *ICCV*, 2003.

[104] Y. Yan, G. Liu, E. Ricci, and N. Sebe. Multi-task linear discriminant analysis for multi-view action recignition. In *ICIP*, 2013.

[105] Y. Yan, E. Ricci, G. Liu, and N. Sebe. Recognizing daily activities from first-person videos with multi-task clustering. In *Asian Conference on Computer Vision*, 2014.

[106] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *ICCV*, 2013.

[107] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *ICCV*, 2013.

[108] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe. Multi-task linear discriminant analysis for multi-view action recognition. *IEEE Transactions on Image Processing*, 23(12):5599–5611, 2014.

[109] Y. Yan, H. Shen, G. Liu, Z. Ma, C. Gao, and N. Sebe. Glocal tells you more: Coupling glocal structural for feature selection with sparsity for image and video classification. *Computer Vision and Image Understanding*, 124:99–109, 2014.

[110] Y. Yan, R. Subramanian, O. Lanz, and N. Sebe. Active transfer learning for multi-view head-pose classification. In *ICPR*, 2012.

[111] Y. Yan, Y. Yang, D. Meng, G. Liu, W. Tong, A. Hauptmann, and N. Sebe. Event oriented dictionary learning for complex event detection. *IEEE Transactions on Image Processing*, 24(6):1867–1878, 2015.

[112] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Oswald Lanz, and Nicu Sebe. No matter where you are: Flexible graph-guided multi-task learn-

ing for multi-view head pose classification under target motion. In *IEEE International Conference on Computer Vision*, 2013.

[113] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Oswald Lanz, and Nicu Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *ICCV*, 2013.

[114] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Gaowen Liu, and Nicu Sebe. Multi-task linear discriminant analysis for multi-view action recognition. *IEEE TIP*, 23(12):5599–5611, 2014.

[115] Yan Yan, Yi Yang, Haoquan Shen, Deyu Meng, Gaowen Liu, Alexander Hauptmann, and Nicu Sebe. Complex event detection via event oriented dictionary learning. In *AAAI*, 2015.

[116] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM MM*, 2007.

[117] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[118] L. Yang, S. Hanneke, and J. Carbonell. A theory of transfer learning with application to actively transfer. *JMLR*, 2012.

[119] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang. Fisher discrimination dictionary learning for sparse representation. In *ICCV*, 2011.

[120] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G. Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *IJCV*, 11, 2014.

[121] Yi Yang, Feiping Nie, Dong Xu, Jiebo Luo, Yueting Zhuang, and Yunhe Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *TPAMI*, 34:723–742, 2012.

[122] Yi Yang, Dong Xu, Feiping Nie, Shuicheng Yan, and Yueting Zhuang. Image clustering using local discriminant models and global integration. *IEEE TIP*, 19(10):2761–2773, 2010.

[123] Victoria Yanulevskaya, Jasper Uijlings, Elia Bruni, Andreza Sartori, Elisa Zamboni, Francesca Bacci, David Melcher, and Nicu Sebe. In the eye of the beholder: Employing statistical analysis and eye tracking for analyzing abstract paintings. In *ACM MM*, 2012.

[124] Y. Yao and G. Dorretto. Boosting for transfer learning with multiple sources. In *CVPR*, 2010.

[125] Chunfeng Yuan, Weiming Hu, Guodong Tian, Shuang Yang, and Haoran Wang. Multi-task sparse learning with beta process prior for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[126] Xiao-Tong Yuan and Shuicheng Yan. Visual classification with multi-task joint sparse representation. In *CVPR*, 2010.

[127] B. Zhang, F.G.B. De Natale, and N. Conci. Recognition of social interactions based on feature selection from visual codebooks. In *ICIP*, 2013.

[128] Jianguang Zhang, Yahong Han, Jinhui Tang, Qinghua Hu, and Jianmin Jiang. What can we learn about motion videos from still images? In *ACM MM*, 2014.

[129] Jianwen Zhang and Changshui Zhang. Multitask bregman clustering. *Neurocomputing*, 74(10):1720–1734, 2011.

[130] Qiang Zhang and Baoxin Li. Discriminative K-SVD for dictionary learning in face recognition. In *CVPR*, 2010.

[131] Y. Zhang and D. Yeung. A convex formulation for learning task relationships in multi-task learning. In *Conference on Uncertainty in Artificial Intelligence*, 2010.

[132] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *Conference on Advances in Neural Information Processing Systems*, 2011.

[133] Jiayu Zhou, Jianhui Chen, and Jieping Ye. Malsar: Multi-task learning via structural regularization. In *Technical Report of Arizona State University*, 2012.