Multiple Tasks are Better than One:

Multi-task Learning and Feature Selection for Head Pose Estimation, Action Recognition and Event Detection

Yan Yan

Advisor: Prof. Nicu Sebe

Department of Information Engineering and Computer Science

University of Trento

A thesis submitted for the degree of

Doctor of Philosophy

April 2014

Acknowledgements

This thesis would not have been possible done without many important people during my Ph.D study. My supervisor, Prof. Nicu Sebe, is not only a good professor but also a good friend. He is always there to help me in my work and life. He has created a perfect M-Hug group. I have really enjoyed working with him. Prof. Elisa Ricci is my mentor who led me into the realm of machine learning and computer vision analysis. Through our collaboration, I have learnt a lot of theoretical knowledge and research skills. Dr. Ramanathan Subramanian, who had supervised me for the first two years of Ph.D study, led me to get acquainted to research. I have learnt a lot from him. Dr. Oswald Lanz, his self-motivation, hard-working spirit and rigorous research attitude has set a good example for me. I am also very pleased to have had the opportunity to work with Prof. Yi Yang. He was always trying to figure out every detail of the research problem and gave me quite insightful advice. Prof. Alexander G. Hauptmann was my supervisor when I visited Carnegie Mellon University. I am impressed by the way he thinks about every research problem and how he leads and unifies the whole team with his personal charm working toward the goal. M-Huggers are my source of happiness. They have made my life in Trento enjoyable. Lastly, I would like to thank my wife Gaowen Liu and my parents for their continuous support for my academic pursuit. I truly appreciate all the people mentioned above and wish everyone happiness and bright future.

Publications

This thesis consists of the following publications:

- Chapter 2:

  -**Y. Yan**, E. Ricci, R. Subramanian, O. Lanz, N. Sebe: "No Matter Where You Are: Flexible Graph-guided Multi-task Learning for Multi-view Head Pose Classification under Target Motion". *IEEE International Conference on Computer Vision*, pages 1177-1184, 2013.

- Chapter 3:

  -**Y. Yan**, E. Ricci, R. Subramanian, G. Liu, N. Sebe: "Multi-task Linear Discriminant Analysis for Multi-view Action Recognition". Pending major revision in *IEEE Transactions on Image Processing*, 2013.

  Idea previously appeared in:

  -**Y. Yan**, G. Liu, E. Ricci, N. Sebe: "Multi-task Linear Discriminant Analysis for Multi-view Action Recognition". *IEEE International Conference on Image Processing*, pages 2842-2846, 2013.

- Chapter 4:

  -**Y. Yan**, H. Shen, G. Liu, Z. Ma, C. Gao, N. Sebe: "GLocal Tells You More: Coupling GLocal Structural for Feature Selection with Sparsity for Image and Video Classification". In *Computer Vision and Image Understanding*, 2014 (In press).

  Idea previously appeared in:

-**Y. Yan**, Z. Xu, G. Liu, Z. Ma, N. Sebe: "GLocal Structural Feature Selection with Sparsity for Multimedia Data Understanding". In *ACM International Conference on Multimedia*, pages 537-540, 2013

- Chapter 5:

  -**Y. Yan**, H. Shen, Y. Yang, D. Meng, A. G. Hauptmann, N. Sebe: "Looking up Your Own Merriam-Webster: Semantic Dictionary Learning for Complex Event Detection". Under review in *International Journal of Computer Vision*, 2014.

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

- **Y. Yan**, E. Ricci, G. Liu, R. Subramanian, N. Sebe: "Clustered Multi-task Linear Discriminant Analysis for View Invariant Color-Depth Action Recognition". In *International Conference on Pattern Recognition*, 2014.

- **Y. Yan**, R. Subramanian, E. Ricci, O. Lanz, N. Sebe: "Evaluating Multi-task Learning for Multi-view Head-pose Classification in Interactive Environments". In *International Conference on Pattern Recognition*, 2014.

- **Y. Yan**, R. Subramanian, O. Lanz, N. Sebe: "Active Transfer Learning for Multi-View Head-Pose Classification". In *IEEE International Conference on Pattern Recognition*, pages 1168-1171, 2012.

- H. Shen, **Y. Yan**, S. Xu, N. Ballas, W. Chen: "Evaluation of Semi-Supervised Learning Method on Action Recognition". In *Multimedia Tools and Applications*, 2014 (In press).

- R. Subramanian, **Y. Yan**, J. Staiano, O. Lanz, N. Sebe: "On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions". *ACM International Conference on Multimodal Interaction*, pages 3-10, 2013.

- G. Liu, **Y. Yan**, C. Gao, W. Tong, A. Hauptmann, N. Sebe: "The Mystery of Faces: Investigating Face Contribution for Multimedia Event Detection". In *ACM International Conference on Multimedia Retrieval*, 2014.

- G. Costante, V. Galieni, **Y. Yan**, M. Fravolini, E. Ricci, P. Valigi: "Explotting Transfer Learning for Personalized View Invariant Gesture Recognition" In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

- S. Cai, C. Wang, **Y. Yan**, Y. Liu: "Analysis of the pencil of conics with double complex contact and its application to camera calibration". In *Journal of Shanghai Jiaotong University (Science)*, 2013(2).

- "CMU Informedia @TRECVID 2013: Surveillance Event Detection (SED)". In *TRECVID Video Retrieval Evaluation Workshop, NIST, Gaithersburg, MD*, 2013.

- "Informedia E-Lamp @ TRECVID 2013 Multimedia Event Detection and Recounting (MED and MER)". In *TRECVID Video Retrieval Evaluation Workshop, NIST, Gaithersburg, MD*, 2013.

The following papers have been submitted and are under review process at the moment of writing this Ph.D thesis:

- "L1-Norm Low-Rank Matrix Factorization by Variational Bayes". submitted to *IEEE Transactions on Neural Networks and Learning Systems*, 2014.

- "It's All About Habits: Exploiting Multi-Task Clustering for Activity of Daily Living Analysis". submitted to *International Conference on Image Processing*, 2014.

- "Minimizing Dataset Bias: Discriminative Multi-task Sparse Coding through Shared Subspace Learning for Image Classification". submitted to *International Conference on Image Processing*, 2014.

- "Event Oriented Dictionary Learning for Complex Event Detection". submitted to *European Conference on Computer Vision*, 2014.

- "Everyday Activity Recognition from First-person Videos with Multi-task Clustering". submitted to *European Conference on Computer Vision*, 2014.

- "Human Interaction In A Nutshell: A Novel Strategy for Analyzing Human Interaction in Unconstrained Videos". submitted to *ACM International Conference on Multimedia*, 2014.
- "Monet Draws Pictures This Way: A Multi-task Dictionary Learning Approach to Discover Art Styles". submitted to *ACM International Conference on Multimedia*, 2014.

Contents

Chapter 1

Introduction

Computer vision is a field that includes methods for acquiring, processing, analyzing, and understanding images and videos and, in general, high-dimensional data from the real world in order to produce numerical or symbolic information. The classical problem in computer vision is that of determining whether or not the image or video data contains some specific object, feature, or activity. This task can normally be solved robustly and without effort by a human, but is still not satisfactorily solved in computer vision for the general case - arbitrary objects in arbitrary situations. The existing methods for dealing with this problem can at best solve it only for specific objects, such as simple geometric objects (*e.g.*, polyhedra), human faces, printed or hand-written characters, or vehicles, and in specific situations, typically described in terms of well-defined illumination, background, and pose of the object relative to the camera.

Machine Learning (ML) and Computer Vision (CV) have been put together during the development of computer vision in the past decade. Nowadays, machine learning is considered as a powerful tool to solve many computer vision problems. Multi-task learning, as one important branch of machine learning, has developed very fast during the past decade. Multi-task learning Evgeniou & Pontil [2004] methods aim to simultaneously learn classification or regression models for a set of related tasks. This typically leads to better models as compared to a learner that does not account for task relationships. The goal of multi-task learning is to improve the performance of learning algorithms by learning classifiers for multiple tasks jointly. This works particularly well if these tasks have some commonality and are generally slightly under-sampled.

In this thesis, we investigate some challenging problems existing in the computer vision area under the multi-task learning framework. Fig.1.1 shows the framework of this thesis. At the first glance of Fig.1.1, probably some questions naturally presented themselves in your mind. *i.e.*, How do we know where a person is looking at from far-field low-resolution cameras? How do we know what each person is doing? And how do we know what this event is? Is it a 'Wedding ceremony' or 'Flash mob gathering'? In the following parts of this thesis, we will answer these questions in detail from the computer vision point of view considering both single and multiple camera setups and from the machine learning point of view, especially under the multi-task learning framework.

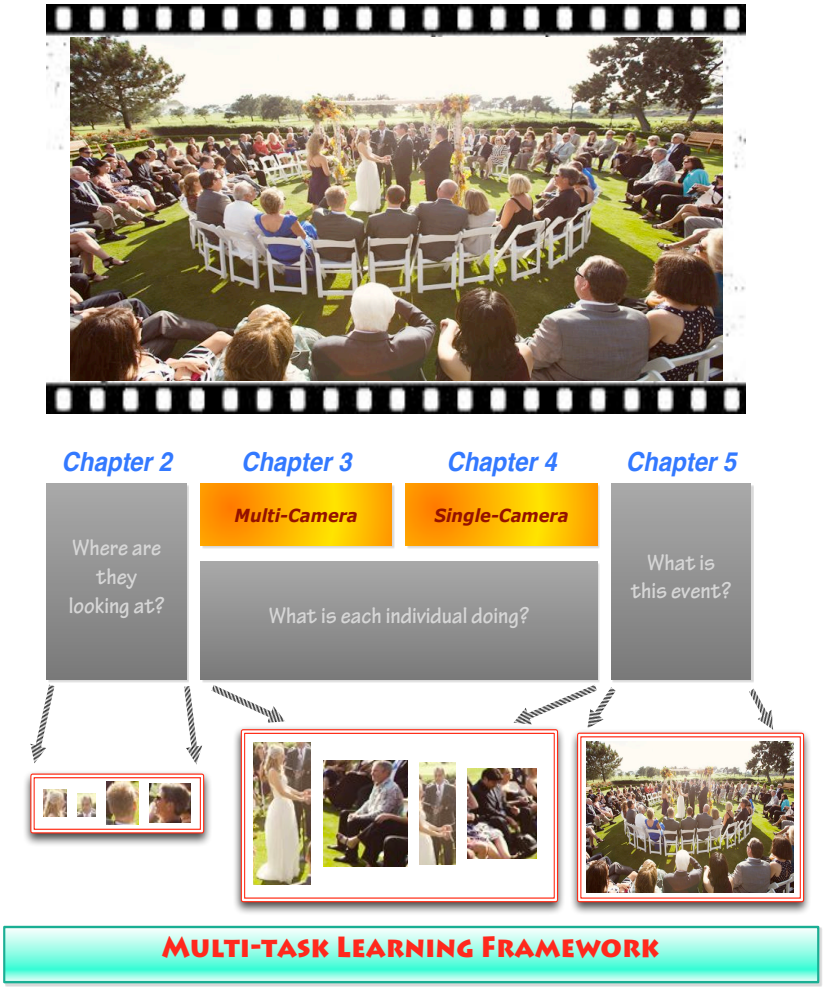


Figure 1.1: The framework of this thesis.

The remainder of this thesis is organized as follows:

- In Chapter 2, we answer the question of 'Where are they looking at?'. Specifically, we propose a novel Multi-task Learning framework (FEGA-MTL) for classifying the head pose of a person who moves freely in an environment monitored by multiple, large field-of-view surveillance cameras.

- In Chapter 3, we answer the question of 'What is each individual doing?' for multi-camera setup. Specifically, we propose Multi-task Linear Discriminant Analysis, a novel multi-task learning framework for multi-view action recognition that allows for the sharing of discriminative Self-Similarity Matrices features among different views.

- In Chapter 4, we answer the question of 'What is each individual doing?' for a single-camera setup. Specifically, we propose a novel feature selection method using a sparse model. Different from the state of the art, our method is built upon $l_{2,p}$-norm and simultaneously considers both the global and local (GLocal) structures of data distribution.

- In Chapter 5, we answer the question of 'What is this event?'. Specifically, we firstly investigate the possibility of automatically selecting semantic meaningful concepts for the event detection task based on both the events-kit text descriptions and the concepts high-level feature descriptions. Then we learn a semantic-oriented dictionary representation for each event based on the selected semantic concepts.

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

- We develop several novel multi-task learning algorithms, *i.e.*, two-graph guided multi-task learning, multi-task linear discriminant analysis, multi-task dictionary learning, which outperform the other state-of-the-art multi-task learning algorithms in the specific computer vision and multimedia problems.

- We provide a novel view from mid-level (headpose and action recognition) computer vision task to high-level (multimedia event detection) multimedia understanding based on multi-task learning approaches.

- We analyse the human action recognition both from single and multiple camera setups under the multi-task learning framework.
- All the proposed algorithms are general frameworks, potentially applicable to many computer vision and pattern recognition problems.

Chapter 2

Flexible Graph-guided Multi-task Learning for Multi-view Head Pose Classification under Target Motion

In this chapter, we propose a novel Multi-Task Learning framework (FEGA-MTL) for classifying the head pose of a person who moves freely in an environment monitored by multiple, large field-of-view surveillance cameras. As the target (person) moves, distortions in facial appearance owing to camera perspective and scale severely impede performance of traditional head pose classification methods. FEGA-MTL operates on a dense uniform spatial grid and learns appearance relationships across partitions as well as partition-specific appearance variations for a given head pose to build region-specific classifiers. Guided by two graphs which *a-priori* model appearance similarity among (i) grid partitions based on camera geometry and (ii) head pose classes, the learner efficiently clusters appearance-wise related grid partitions to derive the optimal partitioning. For pose classification, upon determining the target's position using a person tracker, the appropriate region-specific classifier is invoked. Experiments confirm that FEGA-MTL achieves state-of-the-art classification with few training data.

2.1 Introduction

Head pose estimation and tracking is critical for surveillance and human-behavior understanding, and has been extensively studied for over a decade Murphy-Chutorian & Trivedi [2009]. However, most existing approaches compute the head pose from high resolution images, where facial features are clearly visible. Estimating the head pose from large field-of-view surveillance cameras, where faces are typically captured at 50×50 or lower pixel resolution, has received importance only recently Chen & Odobez [2012]; Orozco *et al.* [2009]; Tosato *et al.* [2010]. Computing the head pose under these conditions is difficult, as faces appear blurred and models employing detailed facial information are ineffective.

Fewer still are head pose estimation methods that utilize information from multiple surveillance cameras. Employing a single camera view is insufficient for studying people's behavior in large environments and multi-view images have been exploited to achieve robust pose estimation Muñoz-Salinas *et al.* [2012]; Rajagopal *et al.* [2012]; Voit & Stiefelhagen [2009]; Yan *et al.* [2012]; Zabulis *et al.* [2009]. However, methods such as Muñoz-Salinas *et al.* [2012]; Voit & Stiefelhagen [2009] estimate pose as a person rotates in place, but is not freely moving around in the environment. The broader goal of this work is to analyze behavior Lepri *et al.* [2012] from head pose cues in unstructured interactive settings (*e.g.* parties), where targets (persons) can move around freely. Therefore, in this paper we consider the problem of *multi-view head pose classification under target motion*.

Fig.2.1(left) illustrates the challenges involved in the considered scenario. The facial appearance of a target with identical 3D head pose but at different positions varies considerably due to perspective and scale. As the target moves, the face can appear larger/smaller and face parts can become occluded/visible due to the target's relative position with respect to the camera. We investigated the effect of appearance change on pose classification using the DPOSE dataset Rajagopal *et al.* [2012], which comprises synchronously recorded images of moving persons from four camera views, associated target positions and head pose annotations. Upon dividing the DPOSE space into four quadrants Q1-Q4, we trained an SVM classifier with HOG Dalal & Triggs [2005] features extracted from the 4-view images corresponding to a particular quadrant. The SVM was then tested with images from each of the quadrants and the task was to assign

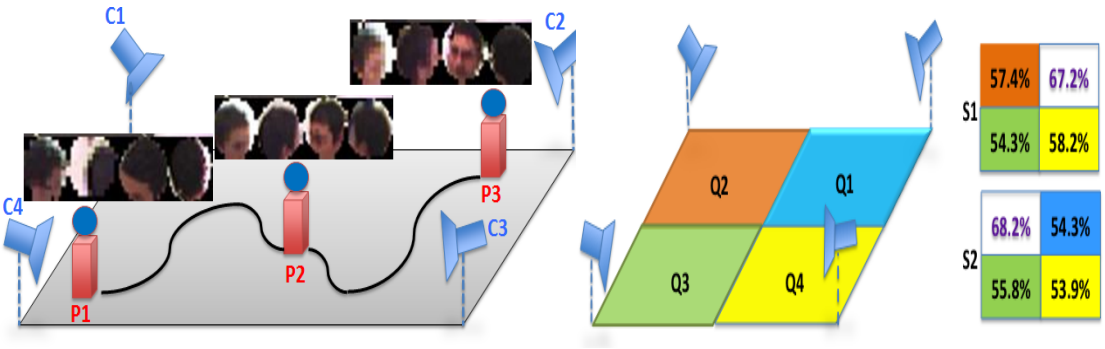


Figure 2.1: (left) Facial appearance change under target motion: for the same 3D head pose, automatically extracted face crops corresponding to camera C1-C4 are shown for target positions P1-P3. (right) Space division: S1, S2 denote classification accuracies when the training images come from the white quadrant (figure is best viewed in color).

head pose to one of eight classes, each denoting a quantized 45^o ($360^o/8$) head-pan. Fig.2.1(right) presents the results. Much lower accuracies were obtained when training and test images came from different quadrants, confirming the adverse impact of position-induced appearance changes on head pose classification.

To address this issue, we propose **FEGA-MTL**, a FlExible GrAph-guided Multi-Task Learning framework for multi-view head pose classification under target motion. Given a set of related tasks, MTL attempts to learn relationships among the tasks as well as task-specific differences. Upon dividing a physical space into a discrete number of planar regions (as in Fig.2.1), we seek to learn the pose-appearance relationship in each region. Analogous with the MTL problem, one can expect some similarity in facial appearance for a given head pose across the regions, and region-specific differences owing to perspective and scale.

FEGA-MTL seeks to simultaneously learn the relationship between facial appearance and head pose across all partitions of a dense uniform 2D spatial grid. Since the facial appearance is likely to be more similar for neighboring regions (as against spatially disjoint partitions), employing a single model to denote the inter-region appearance relationship can lead to *negative transfer*, arising when knowledge sharing has a negative impact on the performance of the learned model. Therefore, we devise a method where appearance-wise related grid clusters (which denote related tasks) are flexibly discovered, and the within-cluster appearance similarity is modeled via the MTL parameters.

Two graphs, which respectively define appearance similarity among (i) grid partitions for a particular head pose given camera geometry, and (ii) head pose classes, guide the learning process to output the optimal spatial partitioning comprising a number of grid clusters and an associated MTL classifier. During the classification stage, upon determining the position corresponding to a test instance using a person tracker, the corresponding region-specific classifier is invoked. Our approach is *flexible* owing to three main reasons: (1) It can work with arbitrary camera setups; (2) The learning algorithm can adaptively deal with multiple feature descriptors having differing discriminative power, and (3) Given the camera geometry and face appearance features, the optimal grid-cluster configuration is automatically discovered using our approach. Experiments confirm that FEGA-MTL outperforms competing head pose classification and MTL approaches.

To summarize, the paper's contributions are: (i) It represents one of the first works to explore multi-view head pose classification under target motion; (ii) To our knowledge, an MTL framework for head pose classification has not been proposed before; (iii) A novel graph-guided approach for simultaneously learning a set of classifiers and their relationships is proposed, and an efficient solver is devised; (iv) We seamlessly connect camera geometry (traditional computer vision) with machine learning for head pose classification through a novel graph modeling strategy; (v) FEGA-MTL is a general framework, potentially applicable to many computer vision and pattern recognition problems.

2.2 Related Work

Head Pose Classification from Low Resolution Faces. Head-pose classification from surveillance images has been investigated in a number of works Benfold & Reid [2011]; Chen & Odobez [2012]; Orozco *et al.* [2009]; Tosato *et al.* [2010]. In Orozco *et al.* [2009], a Kullback-Leibler distance-based facial appearance descriptor is proposed for low resolution images. The array-of-covariances (ARCO) descriptor is introduced in Tosato *et al.* [2010], and is found to be effective for representing faces as it is robust to scale and illumination changes. In Benfold & Reid [2011]; Chen & Odobez [2012], head pose estimation with weak or no supervision is achieved employing motion-based cues and constraints imposed by joint modeling of head and

body pose. However, all these works address single view head pose classification.

Few works estimate head pose fusing information from multiple views Muñoz-Salinas *et al.* [2012]; Rajagopal *et al.* [2012]; Voit & Stiefelhagen [2009]; Zabulis *et al.* [2009]. A particle filter is combined with a neural network for pan/tilt classification in Voit & Stiefelhagen [2009]. A HOG-based confidence measure is also used to determine the relevant views. In Muñoz-Salinas *et al.* [2012], SVMs are employed to calculate a probability distribution for head pose in each view. The results are fused to produce a more precise estimate. Nevertheless, both these works attempt to determine head orientation as a person rotates in place and position-induced appearance variations are not considered.

A weighted distance approach for classifying pose under target motion is proposed in Rajagopal *et al.* [2012]. Upon dividing the space into four quadrants, max-margin distance learning is employed to learn a classifier per region– such a rigid space partitioning scheme will not optimally encode the pose-appearance relationship under motion, with arbitrary camera geometry. In Zabulis *et al.* [2009], head pose under motion is determined by mapping the target's face texture onto a spherical head model, and subsequently locating the face in the unfolded spherical head image. However, many camera views are required to produce an accurate texture map– 9 cameras are used in Zabulis *et al.* [2009]. In contrast, our approach is predominantly image-based, applicable even with few camera views.

Multi-task Learning. MTL methods aim to simultaneously learn classification or regression models for a set of related tasks. This typically leads to better models as compared to a learner that does not account for task relationships. Traditional MTL methods consider a single shared model, assuming that all the tasks are related Argyriou *et al.* [2007]; Evgeniou & Pontil [2004]; Yan *et al.* [2013a]. However, when some of the tasks are unrelated, this may lead to negative transfer. Recently, more sophisticated approaches have been proposed to counter this problem. These methods assume some *a-priori* knowledge (*e.g.* in the form of a graph) defining task dependencies Chen *et al.* [2011] or learn the task relationships simultaneously with task-specific parameters Gong *et al.* [2012]; Jalali *et al.* [2010]; Kang *et al.* [2011]; Zhong & Kwok. [2012]; Zhou *et al.* [2011a]. Among these, the work most similar to ours is Chen *et al.* [2011]. Similar to Chen *et al.* [2011], our algorithm adopts a graph to specify *a-priori* task dependencies. We also overcome the limitations of Chen *et al.* [2011]

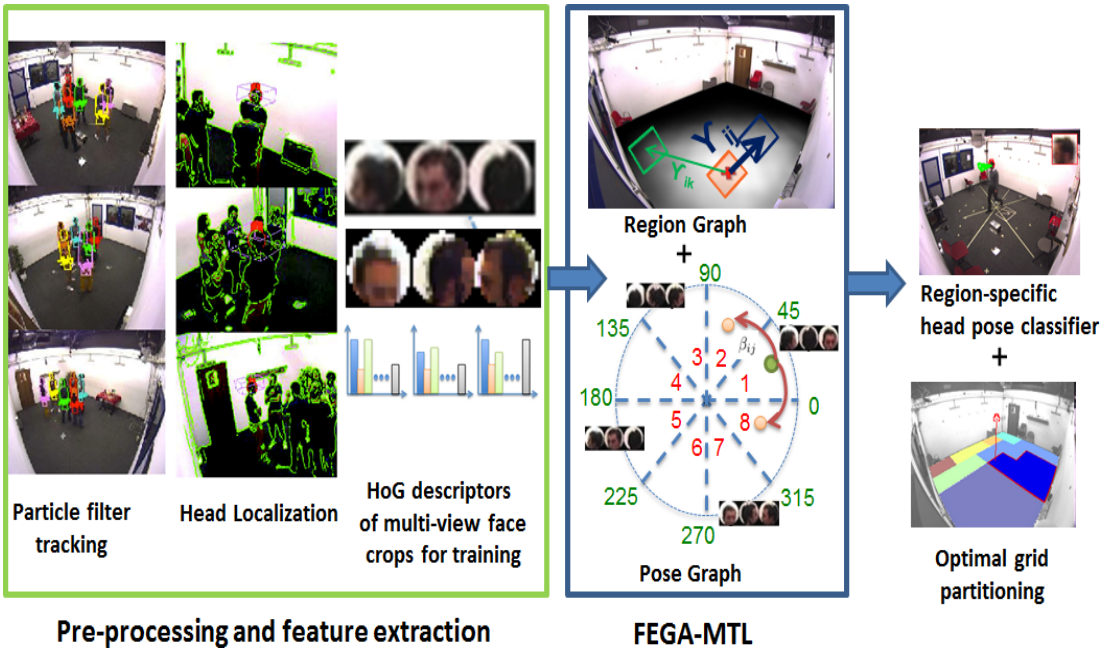


Figure 2.2: Overview of the proposed head pose classification framework assuming three camera views. The region graph and optimal partitioning are as seen from a fourth (camera-less) view. Figure is best viewed in color and under zoom.

as FEGA-MTL automatically discovers task relationships and refines the initial graph structure. For multi-view head pose estimation under motion, the graph structure is very useful as it reflects inter-region facial appearance similarity as derived from the camera geometry.

2.3 Multi-view Head Pose Classification

2.3.1 System Overview

Fig.2.2 presents an overview of our multi-view head pose classification system which consists of three phases: (1) preprocessing and extraction of multi-view face appearance descriptors, (2) learning of head pose-appearance relationships under motion with FEGA-MTL and (3) classification. As we deal with freely moving targets, in the preprocessing stage, a color-based particle filter tracker incorporating multi-view geometry information is employed to reliably localize the target's face and extract multi-view face crops. Also, the tracker allows for determining the target position corresponding to a test instance, so that the appropriate region-based pose classifier can be invoked. Features extracted from the multi-view face appearance images are fed to the FEGA-MTL module for learning region-specific classification parameters.

The learning process is guided by two graphs that respectively model appearance-based task dependencies among grid partitions and head pose classes– (a) the *region graph* quantifies the multi-view facial appearance distortion based on camera geometry, as the target moves from one grid partition to another, and (b) the *pose graph* posits that neighboring head pose classes tend to have more similar facial appearance. FEGA-MTL outputs the pose classification parameters for each grid partition, and the configuration of grid clusters so that the facial appearance for a given head pose is very similar in those partitions constituting a cluster– these grid clusters denote the learnt task relationships given the features and task dependencies. We will now describe each of these modules in detail.

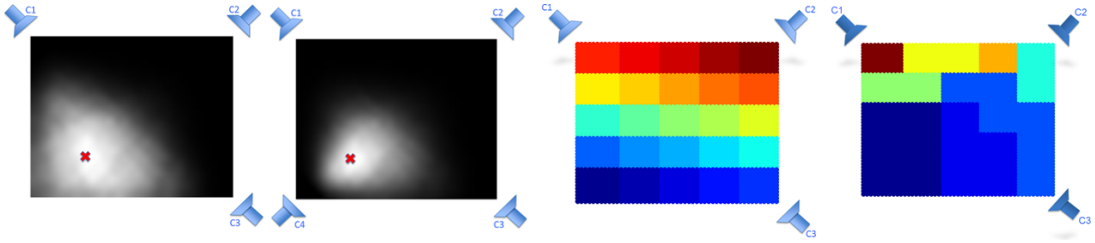


Figure 2.3: (from left to right) Appearance similarity map computed around a point with 3 camera views and 4 camera views, initial grid partitions and learned grid clusters for the 3-camera setup (figure best viewed in color).

2.3.2 Preprocessing

Tracking and Head Localization. A multi-view, color-based particle filter Lanz [2006] is used to compute the 3D body centroid of moving targets. A 30×30×20 cm-sized dense 3D grid (with 1cm resolution) of hypothetic head locations is then placed around the estimated 3D head-position provided by the particle filter[1]. Assuming a spherical model of the head, a contour likelihood is computed for each grid point by projecting a 3D sphere onto each view using camera calibration information. The grid point with the highest likelihood sum is determined as the head location. The tracking and head localization procedures are illustrated in Fig.2.2. The head is then cropped and resized to 20 × 20 pixels in each view.

Feature Extraction. Head crops from the different views are concatenated to generate the multi-view face crops as shown in Fig.2.2, and similar to previous works Benfold & Reid [2011]; Chen & Odobez [2012], we employ HOG descriptors to effectively describe the face appearance for head pose classification. The multi-view face appearance image is divided into non-overlapping 4×4 patches, and a 9-bin histogram is used as the HOG descriptor for each image patch.

2.3.3 Space Partitioning and Graph Modeling

Region Graph Modeling. To apply FEGA-MTL, we initially divide the 2D ground space into a uniform grid with R partitions, as shown in Fig.2.3. We want to learn the pose-appearance relationship in each partition. The algorithm learns from a training

[1]The grid size accounts for the tracker's variance and horizontal and vertical offsets of the head from the body centroid due to pan, tilt and roll.

set $\mathfrak{T}_t = \{(\mathbf{x}_i^t, y_i^t) : i = 1, 2, \ldots, N_t\}$ for each region $t = 1, 2, \ldots, R$, where $\mathbf{x}_i^t \in I\!R^D$ denote D-dimensional feature vectors and $y_i^t \in \{1, 2, \ldots, C\}$ are the head pose labels (C = 8 classes in our setting). One of the graphs guiding the learning process specifies the similarity in appearance for a given head pose across regions based on camera geometry. If grid partitions form the graph nodes, we determine the edge set $\mathcal{E}_1$ and the associated edge weights γ_{mn} quantifying the appearance distortion between $\mathfrak{T}_m$ and $\mathfrak{T}_n$ due to positional change from region m to region n– these edge weights indicate whether knowledge sharing between regions m and n is beneficial or not.

As mentioned earlier, we model the target's head as a sphere. Let Z_k denote the sphere placed at the target's 3D head position p_k, and whose multi-view camera projection yields training image I_k in $\mathfrak{T}_m$. Using camera calibration parameters, one can compute the correspondence between surface points in Z_k and pixels in I_k. Then, we move Z_k to position p_l corresponding to image I_l in $\mathfrak{T}_n$, and determine how many surface points in Z_k are still visible in I_l. The appearance distortion over U camera views due to displacement v from p_k to p_l is defined as $\delta(Z_k, p_k \rightarrow p_l) = \sum_{u=1}^{U} \|v\| + \xi n_0$, where n_0 is the number of surface points in Z_k that are occluded after translation and ξ is a constant that penalizes such occlusion.

The appearance similarity between regions m and n is then computed based on a Gaussian model by considering distortion between all image-pairs associated to $\mathfrak{T}_m$, $\mathfrak{T}_n$ as:

$$\gamma_{mn} = e^{-\frac{\Omega}{N_m N_n \sigma^2}}$$

where $\Omega = \sum_{\forall I_k \in \mathfrak{T}_m, I_l \in \mathfrak{T}_n} [\delta(Z_k, p_k \rightarrow p_l) + \delta(Z_l, p_l \rightarrow p_k)]$, N_m and N_n are number of images in $\mathfrak{T}_m$ and $\mathfrak{T}_n$. $\sigma = 1$ and $\mathcal{E}_1$ is the set of edges for which $\gamma_{mn} \geq 0.1$.

Fig.2.3 depicts the appearance similarity maps for two different camera configurations when the head-sphere at p_k is moved around in space (the projection of p_k on the ground is denoted by the red 'X'). When p_k is close to the camera-less room corner in the 3-camera setup, a number of regions around p_k share a high appearance similarity, implying that pose-appearance relationship can be learnt jointly in these regions. However, the similarity measure decreases sharply as the target moves from p_k towards any of the three cameras, and tends to zero for the upper diagonal half of the room. Also, when a camera is introduced in the fourth room corner, appearance similarity holds only for a smaller portion of space around p_k as compared to the 3-camera case.

Pose Graph Modeling. A second graph guiding the learning process models the fact that facial appearances should be more similar for neighboring pose classes as compared to non-neighboring classes. For example, as shown in Fig.2.2, the facial appearance of exemplars from class 1 should be most similar to exemplars from class 2 and 8. Exploiting this information, a pose graph $\mathcal{E}_2$ is defined with associated edge weights β_{ij} = 1 if i and j correspond to neighboring pose classes c_i, c_j, and β_{ij} = 0 otherwise.

2.4 Flexible Graph-guided MTL

Given a training set $\mathcal{T}_t$, for each task (region) t, we define the matrix $\mathbf{X}_t = [\mathbf{x}_1^t, ..., \mathbf{x}_{N_t}^t]'$, $\mathbf{X}_t \in I\!R^{N_t \times D}$. If $N = \sum_{t=1}^{R} N_t$ denotes the total number of training samples, we also define $\mathbf{X} = [\mathbf{X}_1', \ldots, \mathbf{X}_R']'$, $\mathbf{X} \in I\!R^{N \times D}$ obtained concatenating the matrices $\mathbf{X}_t$ for all the R tasks. In this paper, the notation $(\cdot)'$ indicates the transpose operator. For each training sample, we construct a binary label indicator vector $\mathbf{y}_i^t \in I\!R^{RC}$ as $\mathbf{y}_i^t = [\underbrace{0, 0, ..., 0}_{Task\ 1}, \underbrace{0, 1, ..., 0}_{Task\ 2}, ..., \underbrace{0, 0, ..., 0}_{Task\ R}]$, *i.e.* the position of the non-zero element indicates the task and class membership of the corresponding training sample. A label matrix $\mathbf{Y} \in I\!R^{N \times RC}$ is then obtained concatenating the $\mathbf{y}_i^t$'s for all training samples.

For each region t and pose class c, we propose to learn the region-specific weight vectors for pose classification $\mathbf{w}_{t,c} = \mathbf{s}_{t,c} + \boldsymbol{\theta}_{t,c}$, $\mathbf{w}_{t,c}, \mathbf{s}_{t,c}, \boldsymbol{\theta}_{t,c} \in I\!R^D$. The $\mathbf{s}_{t,c}$ components model the appearance relationships among regions, while $\boldsymbol{\theta}_{t,c}$'s account for region-specific appearance variations. Defining the matrices $\mathbf{S}, \boldsymbol{\Theta} \in I\!R^{D \times RC}$, $\mathbf{S} = [\underbrace{\mathbf{s}_{1,1}, ..., \mathbf{s}_{1,C}}_{Task\ 1}, ..., \underbrace{\mathbf{s}_{R,1}, ..., \mathbf{s}_{R,C}}_{Task\ R}]$, $\boldsymbol{\Theta} = [\underbrace{\boldsymbol{\theta}_{1,1}, ..., \boldsymbol{\theta}_{1,C}}_{Task\ 1}, ..., \underbrace{\boldsymbol{\theta}_{R,1}, ..., \boldsymbol{\theta}_{R,C}}_{Task\ R}]$, we propose to solve the following optimization problem:

$$\min_{\mathbf{S},\boldsymbol{\Theta}} \left\| (\mathbf{Y}'\mathbf{Y})^{-\frac{1}{2}} (\mathbf{Y} - \mathbf{X}(\mathbf{S} + \boldsymbol{\Theta})) \right\|_F^2 + \lambda_s \Omega_s(\mathbf{S}) + \lambda_\theta \Omega_\theta(\boldsymbol{\Theta}) \tag{2.1}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm. The normalization factor $(\mathbf{Y}'\mathbf{Y})^{-1/2}$ compensates for different number of samples per task. The regularization term $\Omega_\theta(\boldsymbol{\Theta}) =$

$\|\boldsymbol{\Theta}\|_F^2$ penalizes large deviation of $\mathbf{s}_{t,c}$ from $\mathbf{w}_{t,c}$, while $\Omega_s(\cdot)$ is defined as follows:

$$\begin{aligned}\Omega_s(\mathbf{S}) =& \|\mathbf{S}\|_F^2 + \lambda_1 \sum_{(i,j)\in\mathcal{E}_1} \gamma_{ij} \|\mathbf{s}_{t_i,c} - \mathbf{s}_{t_j,c}\|_1 \\ &+ \lambda_2 \sum_{(i,j)\in\mathcal{E}_2} \beta_{ij} \|\mathbf{s}_{t,c_i} - \mathbf{s}_{t,c_j}\|_1\end{aligned}$$

where γ_{ij}'s and β_{ij}'s are the appearance similarity-based weights of *region* graph edges $\mathcal{E}_1$ and *pose* graph edges $\mathcal{E}_2$ respectively as described in Sec 2.3.3. The term $\|\mathbf{S}\|_F^2$ regulates model complexity, while the ℓ_1 norm regularizer imposes the weights $\mathbf{s}_{t,c}$ of appearance-wise related regions and neighboring classes to be close together. In particular, region clusters are formed as $\lambda_1 \to \infty$. Importantly, this effect is *feature-specific*– cluster structure varies from feature to feature, and the clustering obtained for the more and less discriminant features can be very different. This is primarily why our method is *flexible*, and the model as well as the proposed optimization strategy benefit from this important effect.

To solve (2.1), we adopt the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) Beck & Teboulle [2009]. FISTA solves optimization problems of the form $\min_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) + r(\boldsymbol{\mu})$, where $f(\boldsymbol{\mu})$ is convex and smooth, $r(\boldsymbol{\mu})$ is convex but non-smooth. Due to its simplicity and scalability, FISTA is a popular tool for solving many convex smooth/non-smooth problems. In each FISTA iteration, a proximal step is computed Beck & Teboulle [2009]:

$$\min_{\boldsymbol{\mu}} \|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_F^2 + \frac{2}{L_k} r(\boldsymbol{\mu}) \tag{2.2}$$

where $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}_k - \frac{1}{L_k}\nabla f(\tilde{\boldsymbol{\mu}}_k)$, $\tilde{\boldsymbol{\mu}}_k$ is the current estimate and L_k is a step-size determined by line search. To apply FISTA to our optimization problem, we define:

$$\begin{aligned}f(\mathbf{S}, \boldsymbol{\Theta}) &= \left\| (\mathbf{Y}'\mathbf{Y})^{-1/2} (\mathbf{Y} - \mathbf{X}(\mathbf{S} + \boldsymbol{\Theta})) \right\|_F^2 \\ r(\mathbf{S}, \boldsymbol{\Theta}) &= \lambda_\theta \|\boldsymbol{\Theta}\|_F^2 + \lambda_s \|\mathbf{S}\|_F^2 + \lambda_s \lambda_1 \sum_{(i,j)\in\mathcal{E}_1} \gamma_{ij} \|\mathbf{s}_{t_i,c} - \mathbf{s}_{t_j,c}\|_1 \\ &\quad + \lambda_s \lambda_2 \sum_{(i,j)\in\mathcal{E}_2} \beta_{ij} \|\mathbf{s}_{t,c_i} - \mathbf{s}_{t,c_j}\|_1\end{aligned}$$

Incorporating the above definition in (2.2) followed by algebraic manipulation, the

Algorithm 1 FEGA-MTL

INPUT: $\mathcal{T}_t, \forall t = 1, \ldots, R$, λ_s, λ_θ, λ_1, λ_2, $\mathbf{E}$
Initialize $\mathbf{S}_0$, $\mathbf{\Theta}_0$, $\alpha_0 = 1$.
OUTER LOOP:
$\alpha_n = \frac{1}{2}(1 + \sqrt{1 + 4\alpha_{n-1}^2})$
{*Update* $\mathbf{\Theta}$}
$\hat{\mathbf{\Theta}} = \mathbf{\Theta}_n - 2\mathbf{X}'(\mathbf{X}\mathbf{\Theta}_n - \mathbf{Y})$
$\mathbf{\Theta}_{n+\frac{1}{2}} = \frac{1}{1+\hat{\lambda}_\theta}\hat{\mathbf{\Theta}}$
$\mathbf{\Theta}_{n+1} = (1 + \frac{\alpha_{n-1}-1}{\alpha_n})\mathbf{\Theta}_{n+\frac{1}{2}} - \frac{\alpha_{n-1}-1}{\alpha_n}\mathbf{\Theta}_n$
{*Update* $\mathbf{S}$}
$\hat{\mathbf{S}} = \mathbf{S}_n - 2\mathbf{X}'(\mathbf{X}\mathbf{S}_n - \mathbf{Y})$
Update $\mathbf{S}_{n+\frac{1}{2}}$ with ADMM as follows:
For each $d = 1 : D$
Initialize $\mathbf{q}^{d,0}$, $\mathbf{a}^{d,0}$, $\mathbf{s}^{d,0}$
Set $\mathbf{M} = \rho\mathbf{E}'\mathbf{E} + (2 + 2\hat{\lambda}_s)\mathbf{I}$
Compute Cholesky factorization of matrix $\mathbf{M}$.
INNER LOOP:
{*Update* $\mathbf{s}$} Solve $\mathbf{M}\mathbf{s}^{d,k+1} = \mathbf{b}^k$
{*Update* $\mathbf{q}$} $\mathbf{q}^{d,k+1} = ST_{\hat{\lambda}_1/\rho}(\mathbf{E}\mathbf{s}^{d,k+1} + \frac{1}{\rho}\mathbf{a}^{d,k})$
{*Update* $\mathbf{a}$} $\mathbf{a}^{d,k+1} = \mathbf{a}^{d,k} + \rho(\mathbf{E}\mathbf{s}^{d,k+1} - \mathbf{q}^{d,k+1})$
Until Convergence
$\mathbf{S}_{n+1} = (1 + \frac{\alpha_{n-1}-1}{\alpha_n})\mathbf{S}_{n+\frac{1}{2}} - \frac{\alpha_{n-1}-1}{\alpha_n}\mathbf{S}_n$
Until Convergence
Output: $\mathbf{W} = \mathbf{S} + \mathbf{\Theta}$

proximal step amounts to solving the following:

$$\min_{\mathbf{S},\mathbf{\Theta}} \quad \left\|\mathbf{S} - \hat{\mathbf{S}}\right\|_F^2 + \left\|\mathbf{\Theta} - \hat{\mathbf{\Theta}}\right\|_F^2 + \hat{\lambda}_1 \sum_{(i,j)\in\mathcal{E}_1} \gamma_{ij}\|\mathbf{s}_{t_i,c} - \mathbf{s}_{t_j,c}\|_1$$
$$+\hat{\lambda}_2 \sum_{(i,j)\in\mathcal{E}_2} \beta_{ij}\|\mathbf{s}_{t,c_i} - \mathbf{s}_{t,c_j}\|_1 + \hat{\lambda}_s \|\mathbf{S}\|_F^2 + \hat{\lambda}_\theta \|\mathbf{\Theta}\|_F^2 \tag{2.3}$$

where $\hat{\lambda}_s = 2\lambda_s/L_k$, $\hat{\lambda}_\theta = 2\lambda_\theta/L_k$ $\hat{\lambda}_1 = 2\lambda_1\lambda_s/L_k$ and $\hat{\lambda}_2 = 2\lambda_2\lambda_s/L_k$, $\hat{\mathbf{\Theta}} = \mathbf{\Theta} - 2\mathbf{X}'(\mathbf{X}\mathbf{\Theta} - \mathbf{Y})$ and $\hat{\mathbf{S}} = \mathbf{S} - 2\mathbf{X}'(\mathbf{X}\mathbf{S} - \mathbf{Y})$. We solve (3.16) by considering $\mathbf{S}$, $\mathbf{\Theta}$ separately, using the procedure described in Algorithm 1. To optimize with respect to $\mathbf{S}$, we devise a novel approach using alternating direction method of multipliers (ADMM) Boyd *et al.* [2011]. In Algorithm 1, $ST_\lambda(x) = \text{sign}(x)\max(|x| - \lambda, 0)$ is a *soft-thresholding* operator and the matrix $\mathbf{E} = \begin{bmatrix} \hat{\lambda}_1\mathbf{E_1} \\ \frac{\hat{\lambda}_2}{\hat{\lambda}_1}\mathbf{E_2} \end{bmatrix}$ is defined considering

the edge-vertex incident matrices $\mathbf{E}_{1\ e=(i,j),h} = \begin{cases} \gamma_{ij},\ i = h \\ -\gamma_{ij},\ j = h \\ 0,\ otherwise \end{cases}$, $\mathbf{E}_1 \in I\!R^{|\mathcal{E}_1| \times RC}$, and $\mathbf{E}_{2\ e=(i,j),h} = \begin{cases} \beta_{ij},\ i = h \\ -\beta_{ij},\ j = h \\ 0,\ otherwise \end{cases}$, $\mathbf{E}_2 \in I\!R^{|\mathcal{E}_2| \times RC}$.

Regarding the computational complexity of Algorithm 1, the main steps in the outer loop are: the update of $\mathbf{\Theta}$ which takes $\mathcal{O}(DR)$ time, the gradient computation taking $\mathcal{O}(N_t DR)$ time and the update of $\mathbf{S}$. The last step is the most computationally expensive as it requires a Cholesky matrix factorization ($\mathcal{O}(R^3)$) for each dimension $d = 1, \ldots, D$. However, the Cholesky factorization is performed only in the outer loop. In the inner loop, each iteration involves solving one linear system ($\mathcal{O}(R^2)$) and a soft-thresholding operation ($\mathcal{O}(|\mathcal{E}_1| + |\mathcal{E}_2|)$).

After the learning phase, the computed weights matrix $\mathbf{W} = \mathbf{S} + \mathbf{\Theta}$ is used for classification. While testing, upon determining the region t associated to a test sample $\mathbf{x}_{test}$ using the person tracker, the corresponding $\mathbf{w}_{t,c}$'s are used to compute the head pose label as $\underset{c=1,\ldots,C}{\arg\max}\ \mathbf{w}'_{t,c}\mathbf{x}_{test}$.

2.5 Experimental Results

In this section, we compare head pose classification results achieved with FEGA-MTL against (i) state-of-the-art head pose estimation methods and (ii) other MTL approaches. We perform our experiments on the DPOSE dataset Rajagopal *et al.* [2012]. To our knowledge, there are no other databases for benchmarking multi-view head pose classification performance under target motion. The CLEAR Stiefelhagen *et al.* [2007] and UcoHead Muñoz-Salinas *et al.* [2012] databases are recorded with targets rotating in-place, while the dataset proposed in Zabulis *et al.* [2009] does not include ground-truth head pose measurements for moving targets. DPOSE comprises over 50000 4-view synchronized images recorded for 16 moving targets, with associated positional and head pose measurements (target positions are computed using the person tracker Lanz [2006]).

As mentioned earlier, the larger goal of this work is to detect interactions in informal gatherings such as *parties*, where we mainly focus on classifying the head-pan

Table 2.1: DPOSE dataset: Head pose classification accuracy. Comparison with state-of-the-art head pose estimation methods.

	4-view				**2-view**			
	Training Set Size/Class/Region				Training Set Size/Class/Region			
	5	10	20	30	5	10	20	30
Single SVM	0.495	0.564	0.65	0.70	0.441	0.486	0.559	0.602
Multiple Region-specific SVMs	0.523	0.571	0.664	0.699	0.446	0.51	0.58	0.618
ℓ_{21} MTL Argyriou *et al.* [2007]	0.589	0.696	0.779	0.795	0.525	0.642	0.724	0.758
Multi-view SVM Muñoz-Salinas *et al.* [2012]	0.544	0.573	0.682	0.713	0.447	0.486	0.565	0.672
ARCO Tosato *et al.* [2010]	0.603	0.70	0.761	0.784	0.529	0.64	0.695	0.739
Single SVM+Warping Rajagopal *et al.* [2012]	0.563	0.644	0.725	0.752	0.466	0.575	0.653	0.687
FEGA-MTL	**0.660**	**0.759**	**0.822**	**0.861**	**0.602**	**0.711**	**0.759**	**0.799**

Table 2.2: DPOSE dataset: Head pose classification accuracy. Comparison with MTL approaches.

	5 training samples/class/region			**10** training samples/class/region		
	2-view	3-view	4-view	2-view	3-view	4-view
Single SVM	0.441	0.494	0.523	0.486	0.549	0.564
ℓ_{21} MTL Argyriou *et al.* [2007]	0.525	0.567	0.589	0.642	0.675	0.696
Flexible Task Clusters MTL Zhong & Kwok. [2012]	0.555	0.598	0.621	0.65	0.681	0.715
Dirty model MTL Jalali *et al.* [2010]	0.546	0.585	0.603	0.655	0.686	0.696
Clustered MTL Zhou *et al.* [2011a]	0.540	0.590	0.619	0.639	0.682	0.711
Robust MTL Gong *et al.* [2012]	0.550	0.580	0.581	0.655	0.689	0.705
FEGA-MTL (region graph only, $\lambda_2 = 0$)	0.581	0.623	0.643	0.677	0.718	0.733
FEGA-MTL (region graph + pose graph)	**0.602**	**0.643**	**0.660**	**0.711**	**0.748**	**0.759**

into one of 8 classes (each denoting a 45° pan range). Since faces are captured at low-resolution by distant, large field-of-view cameras, this task is quite challenging and the state-of-the-art can achieve only about 79% accuracy on the 4-view face images (Table 3.1). We divide DPOSE into mutually exclusive training/validation/test sets. For all methods, regularization parameters are tuned using the validation set, considering values in the interval $[2^{-3}, 2^{-2}, \ldots, 2^{3}]$. We consider an initial, uniformly spaced grid with $R = 25$ regions as shown in Fig.2.3. Our results denote mean classification accuracies obtained from five independent trials, where a randomly chosen training set is employed in each trial.

Table 3.1 presents results comparing FEGA-MTL with competing head pose classification methods. We gradually increase the training set size from 5 to 30 samples/class/region, while the test set comprises images from all regions. As baselines, we consider the recent multi-view approach which probabilistically fuses the

output of multiple SVMs Muñoz-Salinas *et al.* [2012] and the state-of-the-art ARCO classifier Tosato *et al.* [2010] which is shown to be powerful at low resolution (we feed in the 4-view image features to ARCO in order to extend it to the multi-view setup). As shown in the table, both these methods perform poorly with respect to the proposed approach, as they are not designed to account for facial distortions due to scale/perspective changes.

A better strategy in such cases is to compensate for position-induced appearance distortions in some way Rajagopal *et al.* [2012]; Zabulis *et al.* [2009]. The texture-mapping approach presented in Zabulis *et al.* [2009] is shown to be accurate, but many cameras are required for effective texture mapping. Instead, we attempted the warping method proposed in Rajagopal *et al.* [2012], which despite its simplicity is shown to effectively work with few low-resolution views. We implemented a radial basis SVM to determine head pose from the warped 4-view images. Warping is greatly beneficial in the considered scenario as the Single SVM+Warping method significantly outperforms Single SVM.

It is pertinent to point out two differences between our approach and Rajagopal *et al.* [2012]– Rajagopal *et al.* [2012] proposes a pre-defined division of space (the room is divided into 4 quadrants) which is not necessarily optimal for describing the pose-appearance relationship under arbitrary camera geometry. Secondly, task relationships are not considered in Rajagopal *et al.* [2012], and an independent classifier is used for each quadrant. In contrast, FEGA-MTL discovers the optimal configuration of grid clusters that best describes the pose-appearance relationship given camera geometry. Considering task relationships enables FEGA-MTL to achieve higher classification accuracy than a single global classifier (Single SVM), Single SVM+Warping and separate region-specific classifiers that do not consider inter-region appearance relationships (Multiple Region-specific SVMs).

Table 3.1 also presents accuracies obtained with ℓ_{21} MTL Argyriou *et al.* [2007], which assumes all tasks share a common component. As discussed before, negative transfer adversely affects performance of ℓ_{21} MTL, while FEGA-MTL achieves higher accuracy upon flexibly discovering related tasks. We also repeated the experiments employing only two of the four camera views for head pose classification, and while obtained accuracies are expectedly lower in this case, the accuracy trends are still consistent with the 4-view scenario.

Table 3.3 compares classification performance of various MTL methods. The advantage of employing MTL for head pose classification under target motion is obvious since all MTL approaches greatly outperform single SVM. Moreover, having a flexible learning algorithm which is able to infer appearance relationships among regions provides some advantages in terms of classification accuracy. This is confirmed by the fact that in all situations (varying training set sizes and number of camera views) FTC MTL Zhong & Kwok. [2012], Clustered MTL Zhou *et al.* [2011a] and FEGA-MTL achieve superior performance. FEGA-MTL, which independently considers features and employs graphs to explicitly model region and head pose-based appearance relationships, achieves the best performance. The usefulness of modeling both region and pose-based task dependencies through FEGA-MTL is evident on observing the results in Table 3.3. Using the region graph alone is beneficial as such, while employing the region and pose graphs in conjunction produces the best classification performance.

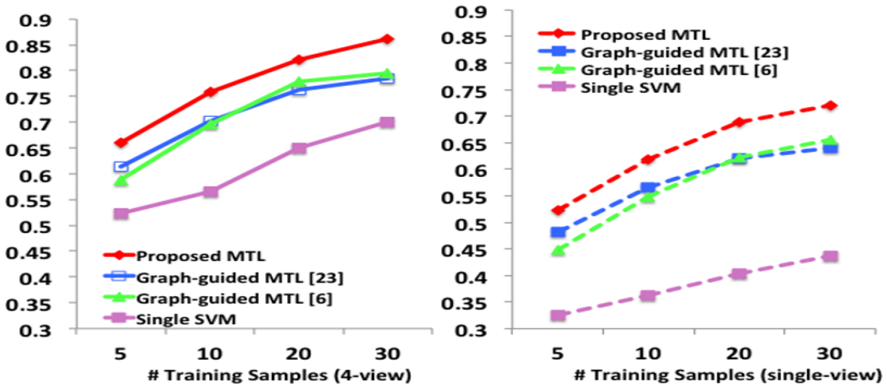


Figure 2.4: Comparison of graph-guided MTL methods: classification accuracies for (left) 4-views and (right) a single view.

Fig.2.3 shows the initial spatial grid and the optimal spatial partitioning learned for a three-camera system with 5 training images/class/region. Clustered regions correspond to identical columns of the task similarity matrix $\mathbf{S}$, *i.e.* two regions t_i and t_j merge if $\mathbf{s}_{t_i,c} = \mathbf{s}_{t_j,c}\ \forall c$. Constrained by the appearance similarity graph weights, spatially adjacent regions tend to cluster together. While regions closer to the camera-less room corner tend to form large clusters, smaller clusters are observed as one moves closer to the cameras owing to larger facial appearance distortions caused by perspective and scale changes. Apart from the region and pose-based appearance similarity

graph weights, facial appearance features also influence the clustering of related regions, and therefore, the computed optimal partitioning.

To further demonstrate the advantages of FEGA-MTL, we compare it with the other graph-guided MTL methods Chen *et al.* [2011]; Zhou *et al.* [2011b]. Fig.2.4 shows that higher accuracy is obtained with our approach for different training set sizes. A main difference between FEGA-MTL and these methods Chen *et al.* [2011]; Zhou *et al.* [2011b] is that they do not decompose $\mathbf{w}_{t,c}$ as $\mathbf{s}_{t,c} + \boldsymbol{\theta}_{t,c}$, and due to the non-consideration of task-specific components $\boldsymbol{\theta}_{t,c}$, they have less flexibility. Moreover, in Zhou *et al.* [2011b] (due to the use of ℓ_2 norm) and Chen *et al.* [2011] (due to smoothing) task-clustering is encouraged but not *enforced*, *i.e.* the $\mathbf{w}_{t,c}$'s corresponding to a cluster are similar but not identical.

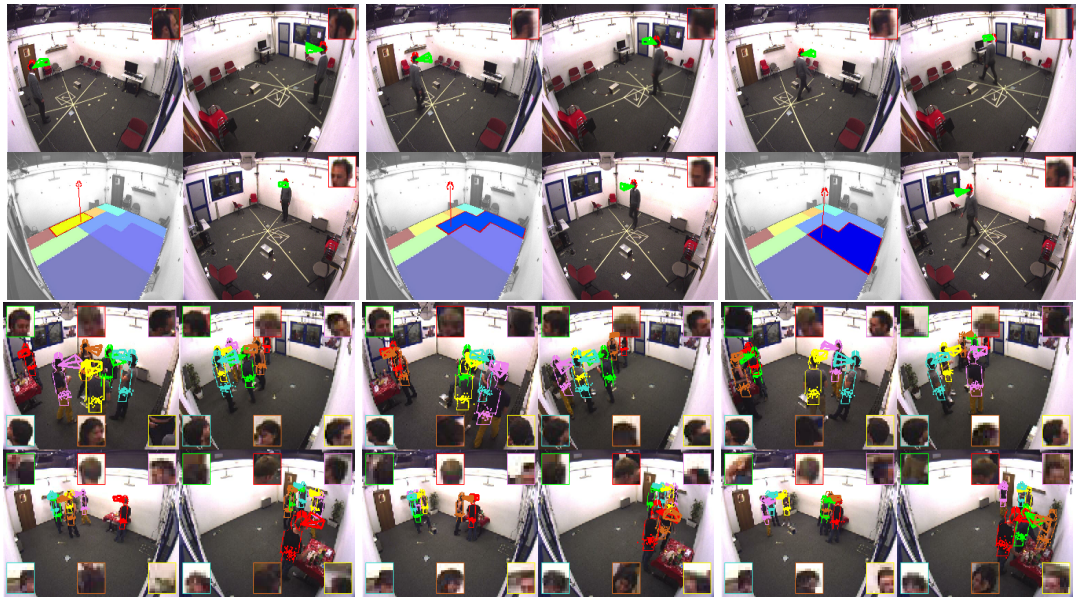


Figure 2.5: (Top) Head pose classification results for a target moving freely within a 3-camera setup are shown two-by-two. The learned clusters, as seen from a fourth view, are shown on the bottom-left inset. Cluster corresponding to the target position (denoted using a stick model) is highlighted. (Bottom) Pose classification results for a party video involving multiple mobile targets (best viewed under zoom.)

Also, it is worth noting that FEGA-MTL can also be used in a single-view setting. However, the use of multiple views is greatly advantageous. Fig.2.4 presents the accuracies obtained with 4-view features against single-view features (mean of the accuracies obtained with each of the four views is considered here). Expectedly, higher

classification accuracy is obtained with the four-view features. The performance gain achieved using FEGA-MTL over an SVM modeling pose-appearance relationship over the entire space is evident, for both single and four-view cases.

Finally, Fig.2.5 shows some qualitative results obtained with FEGA-MTL for single and multiple targets tracked real-time using Lanz [2006]. With multiple targets, identical colors are used to denote the pose direction frustum and face crop rectangle for each target. This scenario is quite challenging, as six targets are interacting naturally and freely moving around in the room.

2.6 Conclusions

We propose a novel graph-guided FEGA-MTL framework for classifying head pose of moving targets from multiple camera views. Starting from a dense 2D spatial grid, two graphs which respectively model appearance similarity among grid partitions and head pose classes guide the learner to output region-specific pose classifiers and the optimal space partitioning. Experiments demonstrate the superiority of FEGA-MTL over competing methods.

Chapter 3

Multi-task Linear Discriminant Analysis for View Invariant Action Recognition

Robust action recognition under viewpoint changes has received considerable attention recently. To this end, Self-Similarity Matrices (SSMs) have been found to be effective view-invariant action descriptors. To enhance the performance of SSM-based methods, we propose **Multi-task LDA**, a novel multi-task learning framework for multi-view action recognition that allows for the sharing of discriminative SSM features among different views (*i.e.* tasks). Inspired by the mathematical connection between multivariate linear regression and Linear Discriminant Analysis (LDA), we model multi-task multi-class LDA as a single optimization problem by choosing an appropriate class indicator matrix. In particular, we propose two variants of graph-guided multi-task LDA: (1) where the graph weights specifying view dependencies are fixed *a priori* and (2) where graph weights are flexibly learnt from the training data. We evaluate the proposed methods extensively on multi-view RGB and RGBD video datasets, and experimental results confirm that the proposed approaches compare favorably with the state-of-the-art.

3.1 Introduction

Human action recognition and understanding from image and video content has attracted considerable attention in computer vision due to its critical role in surveillance, behavior analysis, human-computer interaction, robotics and content-based retrieval. Several solutions have been proposed for action recognition over the years– readers may refer to Poppe [2010]; Weinland *et al.* [2011] for extensive surveys. From the *representation* point of view, the approaches can be mainly classified into methods computing the time evolution of human silhouettes Weinland *et al.* [2010], action cylinders Mahmood *et al.* [2001], space-time shapes Yilmaz & Shah [2005], covariance features Guo *et al.* [2013] and local 3D patch descriptors Laptev *et al.* [2008]. From the *feature extraction* point of view, the various approaches can be categorized into motion-based Efros *et al.* [2003], appearance-based Grundmann *et al.* [2008], space-time volume-based Yilmaz & Shah [2005], space-time interest point-based Lin *et al.* [2010], and Self Similarity Matrices-based Junejo *et al.* [2011].

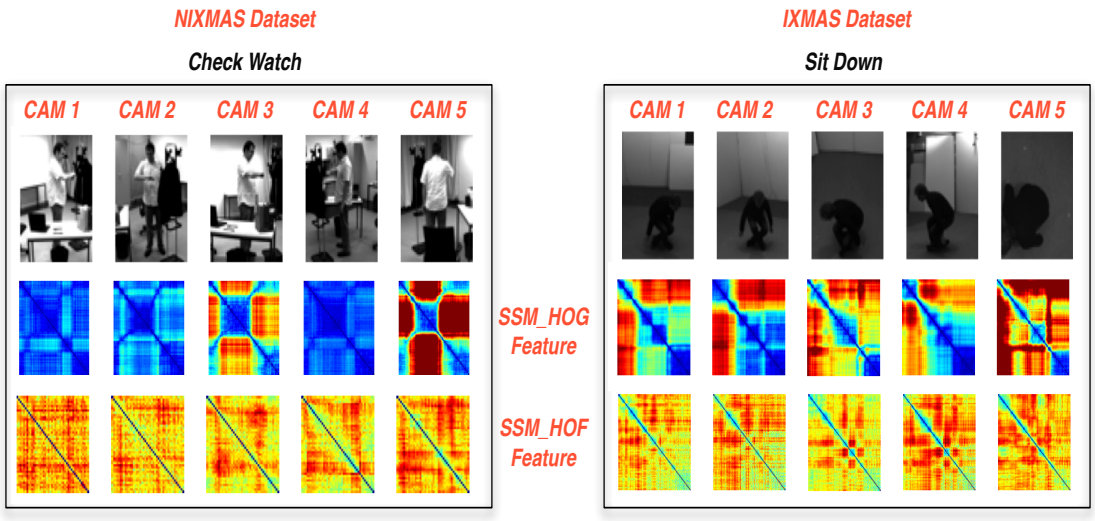


Figure 3.1: Exemplar SSMs computed from Histogram of Oriented Gradients and Histograms of Optical Flows features for the NIXMAS Weinland *et al.* [2010] and IXMAS Weinland *et al.* [2007] datasets. Note the large discrepancy between the HOG-based SSM corresponding to CAM 5 and others for both datasets.

Recently, multi-view action recognition methods have gained in popularity. Since self-occlusion problems can be tackled effectively by employing multiple cameras,

multi-view frameworks can achieve more robust action recognition than monocular methods. However, as actions are typically recognized based on the actor's motion trajectories with respect to the camera viewpoint, viewpoint changes significantly impact action understanding. Therefore, extracting view-invariant information is an important step in multi-view settings but relatively few works have addressed the effect of viewpoint changes on action recognition. Some recent approaches have achieved view-invariant recognition of actions by transferring features across views Farhadi & Tabrizi [2008]; Li & Zickler [2012]; Liu *et al.* [2011] or using view-invariant features Junejo *et al.* [2011]; Li *et al.* [2012]; Rao *et al.* [2002].

A possible methodology for achieving view-invariant action recognition is to compute features which are stable across different viewpoints. Temporal self-similarity matrices (SSMs) Junejo *et al.* [2011], computed from different low-level features such as Histogram of Oriented Gradients (HOG) and Histograms of Optical Flows (HOF), are shown to be robust descriptors for view-invariant action recognition. However, a careful analysis of SSMs reveals that they are also sensitive to large viewpoint-related appearance changes. This effect can be observed in Fig. 3.1, where SSMs corresponding to five views for action sequences from the IXMAS Weinland *et al.* [2007] and NIXMAS Weinland *et al.* [2010] datasets are shown. Although the SSMs associated to all five views share some similarities, it is easy to note that the HoG-based SSM corresponding to the last view (CAM5) is significantly different from the remaining views (CAM1-CAM4) for both datasets.

To arrive at a signature action representation in the presence of large view-related appearance changes, one approach is to find those camera views in which the motion patterns for that action are highly correlated. Multi-task learning (MTL) Argyriou *et al.* [2007]; Evgeniou & Pontil [2004], which simultaneously learns classification/regression models for a set of related tasks, represents an attractive solution to this end. By learning latent relationships between tasks, MTL typically enables the synthesis of models superior to a learner that models each task independently.

In this chapter, we present **Multi-task LDA**, a novel multi-task learning framework to enhance the discriminative power of SSMs for multi-view action recognition, and demonstrate how sharing of features across views (tasks) leads to improved recognition performance. Inspired by the equivalence relationship between multivariate linear regression and linear discriminant analysis (LDA) Ye [2007], we cast multi-task multi-

class LDA as a single optimization problem by choosing an appropriate class indicator matrix, and develop an efficient algorithm to solve it. Also, by defining a graph reflecting prior knowledge on the similarity among different views, the degree of relatedness of the corresponding view features can be controlled using the proposed approach. We describe two variants of graph-guided multi-task LDA and evaluate their performance (1) Multi-task sparse graph-guided LDA, where the graph weights specifying view dependencies are defined *a priori*, and (2) Multi-task flexibile graph-guided LDA where the graph weights are flexibly learnt from (or iteratively refined based on) training features.

Our experiments demonstrate how our methods can be successfully employed for view invariant action recognition, *i.e.* considering the case where images corresponding to the test view are not available in the training data. The obtained results also confirm that sharing features among views is indeed beneficial for multi-view action recognition– our methods outperform competing SSM-based approaches that do not consider task relationships by 10% on the IXMAS dataset. Overall, the proposed approaches achieve efficient recognition of actions on both RGB (video) and RGBD (depth image) data, and compare favorably with the state-of-the-art.

To summarize, the main contributions of this paper are:

- It represents one of the first works to explore a multi-task learning framework for multi-view action recognition. While other recent methods such as Mahasseni & Todorovic [2013] also use MTL, a unique aspect of our framework is that, by effectively combining SSMs descriptors and MTL, it allows for action classification on missing views, for which no examples are available in the training data.

- The proposed approach is shown to be highly effective and achieves improved action recognition performance with respect to other classification methods based on SSM descriptors. While competing works have typically evaluated their algorithms on multi-view video data, we also demonstrate how our framework is applicable to multi-view depth images as in the ACT4^2 Cheng *et al.* [2012] dataset.

- The proposed multi-task LDA framework is novel, and is modeled as a single optimization problem through the use of a class indicator matrix. The described

graph-guided learning algorithms can be generically applied to other computer vision tasks as well.

The remainder of this chapter is organized as follows. Section 2 reviews related work. Section 3 introduces the proposed multi-task linear discriminant analysis and describes the application of our model to multi-view action recognition. Experiments are described in Section 4, while Section 5 concludes this chapter.

3.2 Related Work

In this section, we review related work on multi-view action recognition, linear discriminant analysis and multi-task learning.

3.2.1 Multi-view Action Recognition

Multi-view action recognition has received much attention recently, since a multi-view setup can overcome the problem of self-occlusions and enable more robust action recognition as compared to monocular methods. Both 3D and 2D-based approaches have been proposed for multi-view action recognition as detailed below.

Knowing the 3D scene geometry enables the adaptation of action features from one view to another through the use of geometric transformations. For example, Weinland *et al.* Weinland *et al.* [2007] use 3D occupancy grids synthesized from multiple viewpoints are used to model actions using an exemplar-based HMM. Yen *et al.* Yan *et al.* [2008] employ a 4D action feature model for recognizing actions from arbitrary views. This model encodes shape and motion of actors observed from multiple views, and requires the reconstruction of 3D visual hulls of actors at each time instant. Both approaches lead to computationally intensive algorithms as finding the best match between a 3D model and a 2D observation requires searching over a large model parameter space. Weinland *et al.* Weinland *et al.* [2010] developed a hierarchical classification method based on 3D Histogram of Oriented Gradients (HOG) to represent a test sequence. Robustness to occlusions and viewpoint changes are achieved by combining training data from all viewpoints to train hierarchical classifiers.

A successful approach to tackle the problem of viewpoint-related appearance differences on action recognition in 2D approaches involves the design of view-invariant

features. Rao *et al.* Rao *et al.* [2002] present a view-invariant representation of human actions by capturing changes in the speed and direction of action trajectories using spatio-temporal trajectory curvature. Parameswaran *et al.* Parameswaran & Chellapp [2005] propose to model actions in terms of view-invariant canonical body poses and trajectories in 2D invariance space, which enables representation and recognition of human actions from a generic view-point. Junejo *et al.* Junejo *et al.* [2011] introduced temporal SSMs descriptors as features robust to changes of point of view.

Farhadi and Tabrizi Farhadi & Tabrizi [2008] explicitly address correlations between actions observed from different views. They use a split-based representation to describe clusters of codewords in each view. The transfer of these splits between views is learned from multi-view action sequences. In Farhadi *et al.* [2009], Farhadi *et al.* model view as a latent parameter, and learn features that can discriminate between both views and actions. Liu *et al.* Liu *et al.* [2011] use a bipartite graph to model the relationship between two codebooks generated by k-means clustering of videos acquired for each view. Then, a bipartite partition is used to co-cluster the two view-dependent codebooks into shared visual-word clusters and finally, a codebook composed of these shared clusters is used to encode videos from both views. However, this approach only exploits codebook-to-codebook correspondence at video-level, which cannot guarantee that a pair of videos corresponding to the two views have similar feature representations based on the shared codebook. In addition, it uses a fusion method to combine the prediction outputs of different transferred models, which in turn, requires the clustering of test videos in the target view.

3.2.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is widely used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. This makes LDA a very practical tool for classification and dimensionality reduction. The optimal projection or transformation in classical LDA is obtained by minimizing the within-class distance, and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination.

LDA has been applied successfully in many computer vision applications such

as face recognition Belhumeour *et al.* [1997] or head pose estimation Huang *et al.* [2012b]. Multi-task extensions of LDA have been proposed in Han *et al.* [2010] and Zhang & Yeung [2011]. However in Han *et al.* [2010], the proposed framework is not flexible as no learning on the relationship between tasks is conducted. In Zhang & Yeung [2011] the problem of designing a multi-task LDA algorithm when the different tasks correspond to heterogeneous feature spaces is addressed. However, we do not consider this scenario as it is not appropriate for our application.

3.2.3 Multi-task Learning

Many real-world applications involve related classification/regression tasks. Multi-task learning methods aim to simultaneously learn models for a set of related tasks. By learning tasks in parallel while using a shared representation, improved performance are typically achieved.

Traditional MTL methods consider a single shared model, assuming that all the tasks are related Argyriou *et al.* [2007]; Evgeniou & Pontil [2004]. However, when some of the tasks are unrelated, this may lead to negative transfer and the performance can be even worse than single-task learning. Recently, more sophisticated approaches have been proposed to counter this problem. These methods assume some *a-priori* knowledge (*e.g.* in the form of a graph) defining task dependencies Chen *et al.* [2011] or learning task relationships simultaneously with task-specific parameters Gong *et al.* [2012]. For example, Jalali *et al.* [2010] assume that the data follows a dirty model. Zhou *et al.* [2011a] prove that the clustered MTL approach is equivalent to alternating structure optimization that assumes the tasks sharing a low-dimensional structure. The approach proposed in Zhong & Kwok. [2012] assumes that tasks are clustered, and that clustering structure can be inferred automatically during learning.

Multi-task learning has received considerable attention from the vision community, and has been successfully applied to many problems such as image classification Luo *et al.* [2013], image annotation Tsai *et al.* [2011], visual tracking Zhang *et al.* [2012] and head pose classification under motion Yan *et al.* [2013b]. Recently, an MTL approach to monocular action recognition has been proposed in Zhou *et al.* [2013], where the authors exploit relatedness of action categories to learn latent tasks (motion patterns) shared across actions.

This paper is an extension of previous work presented in Yan *et al.* [2013a], where Multi-task LDA guided by a graph with fixed edge weights is proposed. To our knowledge, multi-view action recognition using multi-task learning has not been considered by other works with the exception of Mahasseni & Todorovic [2013], which is contemporaneous to ours. Our approach and theirs, while focusing both on MTL for multi-view action recognition, are different in the following respects: (1) while Mahasseni & Todorovic [2013] seeks to learn latent action groups, so that within-group feature sharing is allowed but between-group feature sharing is prohibited, we explore learning of latent and discriminative SSM features across views; (2) A part-based action representation is used in Mahasseni & Todorovic [2013], while we use the bag-of-words model for encoding SSM features; (3) A large-margin framework is used for LMTL formulation in Mahasseni & Todorovic [2013], while we propose LDA-based MTL, and (4) While in Mahasseni & Todorovic [2013] the main focus is multi-view action recognition, we also consider the problem of action recognition with missing view, *i.e.* on a novel camera view for which no examples are available in the training set. Furthermore, we show action classification results on the ACT4^2 multi-view depth image dataset, in addition to traditional action video datasets. A description of the proposed Multi-task LDA framework is presented in the following section.

3.3 Multi-task LDA for Multi-view Action Recognition

In this section, we first present an overview of the proposed framework. Then, Self-Similarity Matrix (SSM) descriptors are introduced followed by the analysis of the equivalence between LDA and linear regression. Finally, our multi-task LDA algorithm and its application to the problem of view-independent action recognition are described.

3.3.1 Overview

The proposed approach for view independent action recognition is illustrated in Fig. 3.2. First, different types of low-level features are extracted from videos on a per-frame basis. The type of low-level features extracted depends on the considered sensors: we used Histogram of Oriented Gradients (HOG), Histogram of Optical Flows (HOF)

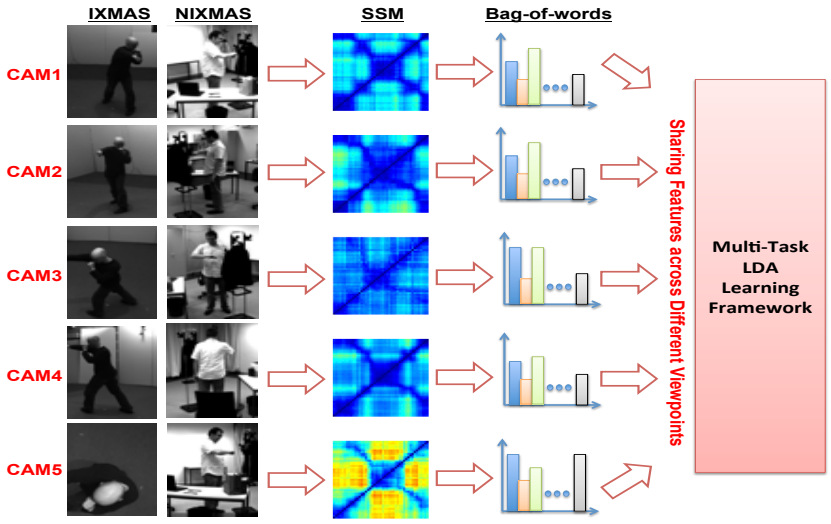


Figure 3.2: Overview of Multi-task LDA-based multi-view action recognition.

and their combination to describe RGB data, while Motion History Images (MHIs) and their variations are adopted to encode information from depth images. Once the SSM descriptors for these low-level features are computed, the standard bag-of-words model is employed for encoding features into histograms. Finally, the proposed multi-task LDA is adopted to induce feature-sharing among the different camera views. Our approach is described in detail in the following subsection.

3.3.2 Self-Similarity Matrix Descriptors

Junejo *et. al.* Junejo *et al.* [2011] introduced SSM descriptors as features robust to viewpoint changes. Given a sequence of images $\mathcal{I} = \{I_1, I_2, ..., I_T\}$, a SSM is a square symmetric matrix:

$$\text{SSM}(\mathcal{I}) = \begin{bmatrix} 0 & e_{12} & e_{13} & \cdots & e_{1T} \\ e_{21} & 0 & e_{23} & \cdots & e_{2T} \\ e_{31} & e_{32} & 0 & \cdots & e_{3T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{T1} & e_{T2} & e_{T3} & \cdots & 0 \end{bmatrix} \tag{3.1}$$

where $e_{ij} = \|\mathbf{f}_i - \mathbf{f}_j\|^2$ is the Euclidean distance between low-level features $\mathbf{f}_i, \mathbf{f}_j$

extracted at frames I_i and I_j respectively. Obviously, as the diagonal corresponds to comparing a frame to itself, it contains all zeros. As low level features in this paper, we use HOG and HOF descriptors on RGB frames, while MHIs are adopted in the case of depth images. The chosen features are described in detail in the experimental section. Once SSMs have been computed (separate SSMs are obtained for each type of low level features), the strategy described in Junejo *et al.* [2011] is adopted for calculating local descriptors. For each point on the diagonal of a single SSM, three local descriptors are computed corresponding to three different diameters in the log-polar domain (28, 42 and 56 frames respectively in diameter). The bag-of-words model is then employed to obtain the final histogram representation of a video clip. A codebook size of 500 words is used in our experiments.

An example of SSMs computed on a sequence extracted from the IXMAS dataset is shown in Fig. 3.1. Obviously, SSMs obtained with different low-level features are different, since each feature captures specific properties of an action. Moreover, SSMs are rather stable over different people performing the same action under multiple viewpoints. However, as observed in the Introduction, SSMs are robust to view changes only up to a certain extent. Therefore, in order to individuate common features from different views, multi-task LDA learning is proposed.

3.3.3 Linear Discriminant Analysis

Linear Discriminant Analysis is a popular technique for dimensionality reduction and classification. We consider a dataset of N samples, $\mathcal{T} = \{(x_i, \ell_i)\}_{i=1}^{N}$, where $x_i \in I\!R^d$ and $\ell_i \in \{1, 2, ..., k\}$ denote respectively the feature vector and the associated class label for the i-th sample, d is the data dimensionality, and k the number of classes. Let $(\cdot)'$ denote the transpose operator. In discriminant analysis Fukunaga [1990], three scatter matrices are defined as follows:

$$S_w = \frac{1}{N} \sum_{j=1}^{k} \sum_{\{x \in \mathcal{T},\ x:\ell=j\}} (x - c_j)(x - c_j)' \tag{3.2}$$

$$S_b = \frac{1}{N} \sum_{j=1}^{k} N_j (c_j - c)(c_j - c)' \tag{3.3}$$

$$S_t = \frac{1}{N}\sum_{i=1}^{N}(x_i - c)(x_i - c)' \tag{3.4}$$

where N_j and c_j denote the number of points and the centroid for the j-th class, while c is the computed centroid of the entire data. It follows from the definition that trace(S_w) and trace(S_b) measure the within-class cohesion and between-class separation respectively. The total scatter matrix is then obtained as $S_t = S_b + S_w$. LDA computes a linear transformation $U \in I\!R^{l\times d}$, mapping the vector $x_i \in I\!R^d$ to a vector $x_i^l \in I\!R^l, x_i^l = Ux_i$, $(l < d)$. In the low dimensional space resulting from the linear transformation U, the scatter matrices become:

$$S_w^l = U'S_wU,\ S_b^l = U'S_bU,\ S_t^l = U'S_tU \tag{3.5}$$

The optimal transformation U^{LDA} is computed solving the following optimation problem Fukunaga [1990]:

$$U^{LDA} = \max_U\ \text{trace}(S_b^l(S_t^l)^{-1}) \tag{3.6}$$

The matrix U^{LDA} is represented by the eigenvectors of $S_t^{-1}S_b$ corresponding to the largest k-1 eigenvalues. In the specific case of a binary-class problems, the optimal transformation Duda *et al.* [2001] is given by:

$$U^{LDA} = S_t^+(c_1 - c_2) \tag{3.7}$$

where c_1 and c_2 are the centroids of the the negative and positive classes respectively.

3.3.4 Linear Regression and LDA

The objective of linear regression is to learn the optimal weight vector $w \in I\!R^d$ such that the function $f(x) = x'w$ can be used to obtain a good estimate of the desired output value ℓ_i, given as input the associated vector x_i. A popular technique for estimating w is the least squares approach, in which the following objective function is minimized:

$$L(w) = \frac{1}{2}\left\|\mathbf{X}'w - \mathbf{y}\right\|^2 \tag{3.8}$$

where $\mathbf{X} = [x_1, x_2, \ldots, x_N]$ is the data matrix and $\mathbf{y} = [\ell_1, \ldots, \ell_N]$ is the vector of class labels. Considering a binary classification problem and assuming that both the data vectors and labels have been centered (*i.e.* $\sum_i^N x_i = 0$ and $\sum_i^N \ell_i = 0$), it follows that $\ell_i \in \{-2N_2/N, 2N_1/N\}$ where N_1 and N_2 denote the number of samples from the negative and positive classes respectively. The optimal w is given by $w = (\mathbf{X}\mathbf{X}')^+\mathbf{X}\mathbf{y}$ Hastie *et al.* [2001]. Noticing that $\mathbf{X}\mathbf{X}' = NS_t$ and $\mathbf{X}\mathbf{y} = \frac{2N_1N_2}{N}(c_1 - c_2)$ it follows that:

$$w = \frac{2N_1N_2}{N^2} S_t^+ (c_1 - c_2) = \frac{2N_1N_2}{N^2} U^{LDA} \tag{3.9}$$

where U^{LDA} is the optimal solution to LDA in (3.7). Hence linear regression with the class labels as output values is equivalent to LDA, as the projection in LDA is invariant to scaling Duda *et al.* [2001].

Recently, similar results have been proven for multi-class LDA Ye [2007] showing that it is equivalent to multivariate linear regression if an indicator matrix $\bar{\mathbf{Y}} \in I\!R^{N \times k}$ is defined as follows:

$$(\bar{\mathbf{Y}})_{ij} = \begin{cases} \sqrt{\frac{N}{N_j}} - \sqrt{\frac{N_j}{N}} & if \;\; \ell_i = j \\ -\sqrt{\frac{N_j}{N}} & otherwise \end{cases} \tag{3.10}$$

where $(\cdot)_{ij}$ is the element in the i-th row, j-th column of the matrix. The optimal projection matrix $\mathbf{W} \in I\!R^{k \times d}$ is obtained by solving the following optimization problem:

$$\min_{\mathbf{W}} \quad \frac{1}{2} \left\| \mathbf{X}'\mathbf{W} - \bar{\mathbf{Y}} \right\|_F^2 \tag{3.11}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. Further details about this equivalence can be found in Ye [2007].

3.3.5 Multi-task Linear Discriminant Analysis

In this paper, an extension of multiclass LDA Ye [2007] to a multi-task learning setting is proposed. We consider a set of R related tasks. Each task is a multi-class classification problem with C categories. For each task $t = 1, 2, \ldots, R$, a training set $\mathcal{T}_t = \{(x_n^t, \ell_n^t)\}_{n=1}^{N_t}$ is given, where $x_n^t \in I\!R^d$ is d-dimensional feature vector, and $\ell_n^t \in \{1, 2, \ldots, C\}$ is the label indicating the class membership. For each task t we

define the matrices $\mathbf{x}_t \in I\!R^{N_t \times d}$, $\mathbf{x}_t = [x_1^t, ..., x_{N_t}^t]'$, $\mathbf{y}_t \in I\!R^{N_t \times C}$, $\mathbf{y}_t = [\ell_1^t, ..., \ell_{N_t}^t]'$ which is the class indicator matrix defined as follows:

$$(\mathbf{y}_t)_{ij} = \begin{cases} \sqrt{\frac{N_t}{N_{t,j}}} - \sqrt{\frac{N_{t,j}}{N_t}} & if \;\; \ell_i^t = j \\ -\sqrt{\frac{N_{t,j}}{N_t}} & otherwise \end{cases} \tag{3.12}$$

where $N_{t,j}$ is the sample size of the j-th class in the t-th task, $N_t = \sum_{j=1}^{C} N_{t,j}$ is total training samples of all classes in the t-th task. Concatenating $\mathbf{x}_t$ and $\mathbf{y}_t$ of all the R tasks, the matrices $\mathbf{X} = [\mathbf{x}_1', \ldots, \mathbf{x}_R']'$, $\mathbf{X} \in I\!R^{N \times d}$ and $\mathbf{Y} = [\mathbf{y}_1', \ldots, \mathbf{y}_R']'$, $\mathbf{Y} \in I\!R^{N \times CR}$ are obtained, where $N = \sum_{t=1}^{R} N_t$. In this paper, we propose to learn a global weight matrix $\mathbf{U} = [\mathbf{u}_1', \ldots, \mathbf{u}_R']', \mathbf{U} \in I\!R^{d \times CR}$ by solving the following optimization problem:

$$\min_{\mathbf{U}} \quad \Lambda(\mathbf{U}) = \frac{1}{2} \left\| (\mathbf{Y} - \mathbf{X}\mathbf{U}) \right\|_F^2 + \Omega(\mathbf{U}) \tag{3.13}$$

where $\Omega(\mathbf{U})$ is an appropriate regularization term. We propose two variant approaches to multi-task LDA, corresponding to different regularization terms $\Omega(\mathbf{U})$. We present them in the following subsections.

3.3.6 Multi-task Sparse Graph Guided LDA

The first approach we propose consists of the following optimization problem:

$$\min_{\mathbf{U}} \quad \frac{1}{2} \left\| (\mathbf{Y} - \mathbf{X}\mathbf{U}) \right\|_F^2 + \lambda_1 \left\| \mathbf{M}\mathbf{U}' \right\|_F^2 + \lambda_2 \left\| \mathbf{U} \right\|_1 \tag{3.14}$$

where $\| \cdot \|_1$ denote the L_1 norm. The matrix $\mathbf{M}$ is an edge-vertex incident matrix, $\mathbf{M} \in I\!R^{|\mathcal{E}| \times CR}$, where $|\mathcal{E}|$ denotes graph set cardinality and:

$$(\mathbf{M})_{q=(i,j),h} = \begin{cases} \gamma_{ij} & if \;\; i = h \\ -\gamma_{ij} & if \;\; j = h \\ 0 & otherwise \end{cases}$$

Here, $\gamma_{ij} = (\sum_{i \neq j} \|SSM_i - SSM_j\|_2)^{-1}$, *i.e.*, γ_{ij} is set by calculating the inverse of the normalized euclidean distance of SSMs descriptors between two different views (tasks) for the same action/class, averaged on the training data. γ_{ij} is normalized into the interval $[0, 1]$ and a large γ_{ij} indicates high similarity of specific action/class between views.

The proposed objective function has three effects. All tasks are related thanks to the graph regularization term, and therefore knowledge from one task can be utilized by the other tasks. Prior knowledge about the required level of sharing feature is embedded in the learning framework through γ_{ij}. Sparsity enforced in the learning process provides a beneficial effect on feature selection, and de-emphasizes the contribution of less discriminative features.

To compensate for the different number of samples per class, we also propose to integrate a term $(\mathbf{Y}\mathbf{Y}')^{-1/2}$ as a normalization factor in the loss function. Noticing that the resulting objective function in (3.14) can be decomposed into two parts, *i.e.* a smooth term $\Pi(\cdot)$ and a non smooth term $\Omega(\cdot)$,

$$\Pi(\mathbf{U}) = \frac{1}{2} \left\| (\mathbf{Y}\mathbf{Y}')^{-1/2}(\mathbf{Y} - \mathbf{X}\mathbf{U}) \right\|_F^2 + \lambda_1 \left\| \mathbf{M}\mathbf{U}' \right\|_F^2, \qquad \Omega(\mathbf{U}) = \lambda_2 \left\| \mathbf{U} \right\|_1$$

we adopt the well-known accelerated gradient method Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) Beck & Teboulle [2009] to solve (3.14) with respect to $\mathbf{U}$ as described in Algorithm 1. FISTA solves the optimization problems in the form $\min_{\mathbf{U}} \Pi(\mathbf{U}) + \Omega(\mathbf{U})$, where $\Pi(\mathbf{U})$ is convex and smooth, $\Omega(\mathbf{U})$ is convex but non-smooth. Due to its simplicity and scalability, FISTA has become a popular tool. In each FISTA iteration, a proximal step is computed Beck & Teboulle [2009]:

$$\min_{\mathbf{U}} \left\| \mathbf{U} - \hat{\mathbf{U}} \right\|_F^2 + \frac{2}{L_k} \Omega(\mathbf{U})$$

where $\hat{\mathbf{U}} = \tilde{\mathbf{U}}_k - \frac{1}{L_k} \nabla \Pi(\tilde{\mathbf{U}}_k)$, $\tilde{\mathbf{U}}_k$ is the current iterate and L_k is a stepsize determined by line search. To solve the proximal step, the soft-thresholding operator $\Sigma_\lambda(x) = \text{sign}(x) \max(|x| - \lambda, 0)$ is adopted Boyd *et al.* [2011]. In Algorithm 1 L_k is the line search step length and $\hat{\lambda}_1 = 2\lambda_1 / L_k$, $\hat{\lambda}_2 = 2\lambda_2 / L_k$.

Algorithm 2 Multi-task Sparse Graph Guided LDA

Input: $\mathcal{T}_t = \{(x_n^t, \ell_n^t)\}_{n=1}^{N_t}, \forall t = 1, \ldots, R$, λ_1, λ_2, $\mathbf{M}$

Initialize $\mathbf{U}_0$, $\alpha_0 = 1$.
LOOP:
$\alpha_k = \frac{1}{2}(1 + \sqrt{1 + 4\alpha_{k-1}^2})$
$\hat{\mathbf{U}} = \mathbf{U}_k - \frac{2}{L_k}[\mathbf{X}'(\mathbf{Y}\mathbf{Y}')^{-1}(\mathbf{X}\mathbf{U}_k - \mathbf{Y}) + \hat{\lambda}_1 \mathbf{U}_k \mathbf{M}'\mathbf{M}]$
Solve $\mathbf{U}_{k+\frac{1}{2}} \leftarrow \min_{\mathbf{U}} \left\|\mathbf{U} - \hat{\mathbf{U}}\right\|_F^2 + \hat{\lambda}_2 \|\mathbf{U}\|_1$ using $\Sigma_\lambda(x) = \text{sign}(x)\max(|x| - \lambda, 0)$.
$\mathbf{U}_{k+1} = (1 + \frac{\alpha_{k-1}-1}{\alpha_k})\mathbf{U}_{k+\frac{1}{2}} - \frac{\alpha_{k-1}-1}{\alpha_k}\mathbf{U}_k$
Until Convergence

Output: $\mathbf{U}$

3.3.7 Multi-task Flexible Graph Guided LDA

The second multi-task LDA approach we propose also develops from (3.13), but instead of fixing the graph weights modeling task dependencies as in (3.14), learns them them from training data. We define $\mathbf{U} = \mathbf{C} + \mathbf{S}$, $\mathbf{C}, \mathbf{S} \in I\!R^{d \times CR}$, *i.e.*, we consider the weight matrix as the matrix obtained summing two terms, the matrix $\mathbf{C}$ modeling common features among tasks, and the matrix $\mathbf{S}$ which accounts for task-specific features. We formulate the following optimization problem derived from (3.13):

$$\min_{\mathbf{C},\mathbf{S}} \|\mathbf{Y} - \mathbf{X}(\mathbf{C} + \mathbf{S})\|_F^2 + \lambda\Omega(\mathbf{C}, \mathbf{S}) \tag{3.15}$$

where $\mathbf{C} = [\mathbf{c}_1', \ldots, \mathbf{c}_R']'$, $\mathbf{S} = [\mathbf{s}_1', \ldots, \mathbf{s}_R']'$ and $\Omega(\cdot)$ is an appropriate regularization term defined as:

$$\Omega(\mathbf{C}, \mathbf{S}) = \|\mathbf{C}\|_F^2 + \|\mathbf{S}\|_F^2 + \lambda_c \|\mathbf{M}\mathbf{C}'\|_1$$

To define the matrix $\mathbf{M}$, we consider the graph structure described in the previous section. In the regularization term, $\|\mathbf{C}\|_F^2$ regulates model complexity, $\|\mathbf{S}\|_F^2$ penalizes large deviation of $\mathbf{c}_t$ from $\mathbf{u}_t$ $\forall t$. The L_1 norm regularizer imposes the weights $\mathbf{c}_t$ of related tasks to be close together and become identical as $\lambda_c \to \infty$, leading to task clusters.

Algorithm 3 Multi-task Flexible Graph Guided LDA

INPUT: $\mathcal{T}_t = \{(x_n^t, \ell_n^t)\}_{n=1}^{N_t}, \forall t = 1, \ldots, R$, λ, λ_c, $\mathbf{M}$.

Initialize $\mathbf{C}_0$, $\mathbf{S}_0$, $\alpha_0 = 1$.
OUTER LOOP:
$\alpha_k = \frac{1}{2}(1 + \sqrt{1 + 4\alpha_{k-1}^2})$
$\hat{\mathbf{C}} = \mathbf{C}_k - 2\mathbf{X}^T(\mathbf{X}\mathbf{C}_k - \mathbf{Y})$
FOR $i = 1 : D$
Initialize $\mathbf{q}^{i,0}$, $\mathbf{z}^{i,0}$, $\mathbf{c}^{i,0}$
Compute $\mathbf{A} = \rho\mathbf{M}^T\mathbf{M} + (2 + 2\hat{\lambda}_c)\mathbf{I}$.
Compute Cholesky factorization of matrix $\mathbf{A}$.
INNER LOOP:
$\mathbf{b}^n = \rho\mathbf{M}^T\mathbf{q}^{i,n} - \mathbf{M}^T\mathbf{z}^{i,n} + 2\hat{\mathbf{c}}^i$
Solve $\mathbf{A}\mathbf{c}^{i,n+1} = \mathbf{b}^n$
$\mathbf{q}^{i,n+1} = \Sigma_{\hat{\lambda}_1/\rho}(\mathbf{M}\mathbf{c}^{i,n+1} + \frac{1}{\rho}\mathbf{z}^{i,n})$
$\mathbf{z}^{i,n+1} = \mathbf{z}^{i,n} + \rho(\mathbf{M}\mathbf{c}^{i,n+1} - \mathbf{q}^{i,n+1})$
Until Convergence
END FOR
$\mathbf{C}_{k+1} = (1 + \frac{\alpha_{k-1}-1}{\alpha_k})\mathbf{C}_{k+\frac{1}{2}} - \frac{\alpha_{k-1}-1}{\alpha_k}\mathbf{C}_k$
$\hat{\mathbf{S}} = \mathbf{S}_k - 2\mathbf{X}^T(\mathbf{X}\mathbf{S}_k - \mathbf{Y})$
$\mathbf{S}_{k+\frac{1}{2}} = \frac{1}{1+\hat{\lambda}}\hat{\mathbf{S}}$
$\mathbf{S}_{k+1} = (1 + \frac{\alpha_{k-1}-1}{\alpha_n})\mathbf{S}_{k+\frac{1}{2}} - \frac{\alpha_{k-1}-1}{\alpha_k}\mathbf{S}_k$
Until Convergence
Output: $\mathbf{U} = \mathbf{C} + \mathbf{S}$

To apply the FISTA approach to our optimization problem we define:

$$\Pi(\mathbf{C}, \mathbf{S}) = \|\mathbf{Y} - \mathbf{X}(\mathbf{C} + \mathbf{S})\|_F^2$$
$$\Omega(\mathbf{C}, \mathbf{S}) = \lambda\|\mathbf{S}\|_F^2 + \lambda\|\mathbf{C}\|_F^2 + \lambda\lambda_c \|\mathbf{M}\mathbf{C}'\|_1$$

The proximal step amounts into solving the following:

$$\min_{\mathbf{C},\mathbf{S}} \quad \left\|\mathbf{C} - \hat{\mathbf{C}}\right\|_F^2 + \left\|\mathbf{S} - \hat{\mathbf{S}}\right\|_F^2 + \hat{\lambda}_c \|\mathbf{M}\mathbf{C}'\|_1 + \hat{\lambda}\|\mathbf{C}\|_F^2 + \hat{\lambda}\|\mathbf{S}\|_F^2 \tag{3.16}$$

where $\hat{\lambda} = 2\lambda/L_k$ and $\hat{\lambda}_c = 2\lambda\lambda_c/L_k$, $\hat{\mathbf{S}} = \mathbf{S} - 2\mathbf{X}^T(\mathbf{X}\mathbf{S} - \mathbf{Y})$ and $\hat{\mathbf{C}} = \mathbf{C} -$

$2\mathbf{X}^T(\mathbf{XC} - \mathbf{Y})$. To solve (3.16), we consider $\mathbf{C}$, $\mathbf{S}$ separately. While solving with respect to $\mathbf{S}$ is straightforward, solving with respect to $\mathbf{C}$ is more challenging due to the presence of the L_1 norm. However, since in our approach each feature dimension is considered independently, the update of the weight vectors $\mathbf{C}$ can be made very efficient by solving d separate optimization problems (one for each row $\mathbf{c}^i$ of the matrix $\mathbf{C}$) as:

$$\min_{\mathbf{c}^i} \left\|\mathbf{c}^i - \hat{\mathbf{c}}^i\right\|_2^2 + \hat{\lambda}_c \left\|\mathbf{M}\mathbf{c}^i\right\|_1 + \hat{\lambda} \left\|\mathbf{c}^i\right\|_2^2$$

In this paper we propose to apply the augmented Lagrangian multipliers approach Boyd *et al.* [2011], and consider the equivalent constrained optimization problem (in the following the superscripts are removed for sake of clarity):

$$\begin{aligned} &\min_{\mathbf{c},\mathbf{q}} \|\mathbf{c} - \hat{\mathbf{c}}\|_2^2 + \hat{\lambda}_c \|\mathbf{q}\|_1 + \hat{\lambda} \|\mathbf{c}\|_2^2 \\ &\text{s.t.} \quad \mathbf{M}\mathbf{c} - \mathbf{q} = 0 \end{aligned} \tag{3.17}$$

Defining with $\mathbf{z}$ being the vector of augmented Lagrangian multipliers and ρ being the dual update step-length, the associated Lagrangian is:

$$\begin{aligned} L_\rho(\mathbf{c}, \mathbf{q}, \mathbf{z}) = \|\mathbf{c} - \hat{\mathbf{c}}\|_2^2 + \hat{\lambda}_c \|\mathbf{q}\|_1 + \hat{\lambda} \|\mathbf{c}\|_2^2 \\ + \mathbf{z}^T(\mathbf{M}\mathbf{c} - \mathbf{q}) + \tfrac{\rho}{2} \|\mathbf{M}\mathbf{c} - \mathbf{q}\|_2^2 \end{aligned} \tag{3.18}$$

Three steps are alternated, corresponding to solving (3.18) with respect to the three sets of variables $\mathbf{c}$, $\mathbf{q}$ and $\mathbf{z}$. Solving with respect to $\mathbf{c}$, with $\mathbf{q}, \mathbf{z}$ fixed, implies solving a linear system $\mathbf{A}\mathbf{c}^{k+1} = \mathbf{b}^k$ where $\mathbf{A} = \rho\mathbf{M}^T\mathbf{M} + (2 + 2\hat{\lambda}_c)\mathbf{I}$ and $\mathbf{b}^k = \rho\mathbf{M}^T\mathbf{q}^k - \mathbf{M}^T\mathbf{z}^k + 2\hat{\mathbf{c}}$. Cholesky factorization can be used to decompose $\mathbf{A}$ and solve the linear system efficiently. Solving with respect to $\mathbf{q}$ has a closed form solution obtained applying the soft-thresholding operator. The update step corresponding to solving with respect to $\mathbf{z}$ is straightforward. The procedure is outlined in Algorithm 3.

3.4 Experimental Results

In this section, the performance of the proposed approaches is assessed on three publicly available multi-view action recognition datasets, namely IXMAS Weinland *et al.* [2007], NIXMAS Weinland *et al.* [2010] and ACT4^2 Cheng *et al.* [2012].

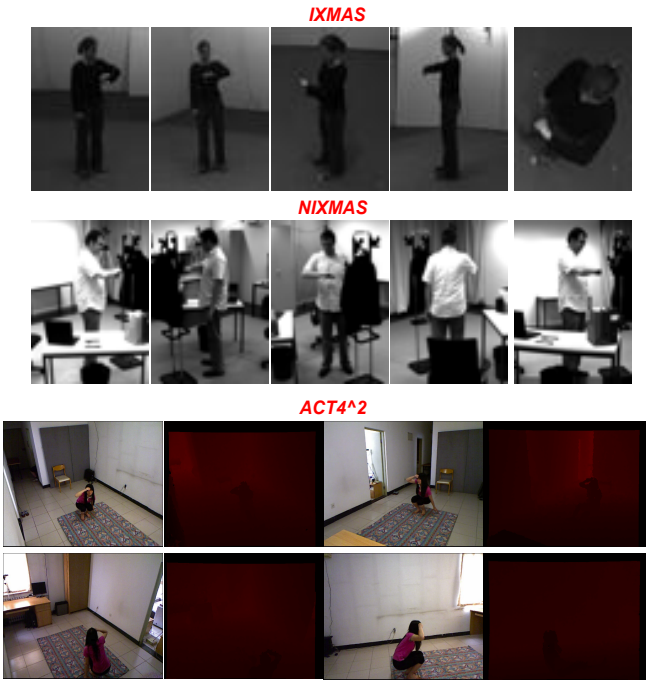


Figure 3.3: Sample frames extracted from the three considered datasets. A majority of the action sequences in the NIXMAS dataset involve occlusions due to other scene objects. Figure is best viewed in color.

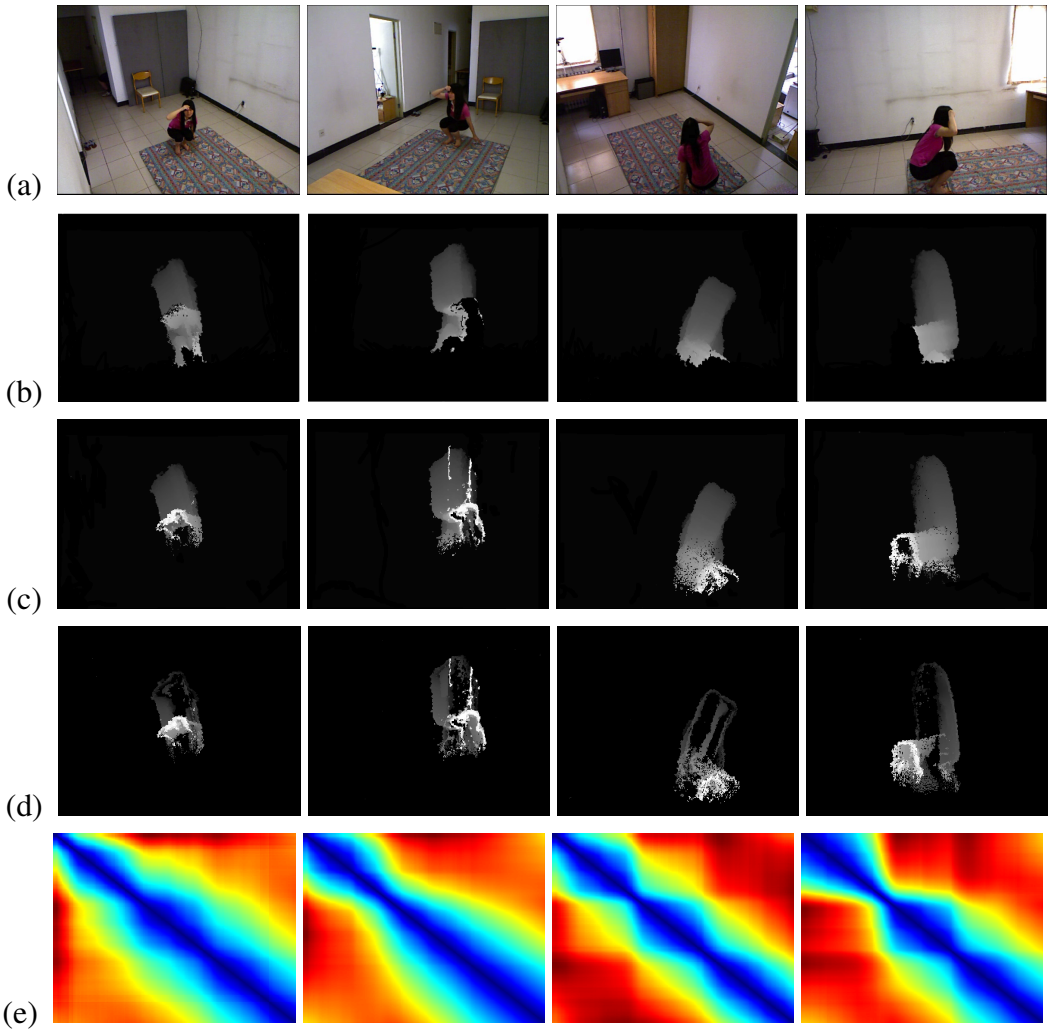


Figure 3.4: ACT4^2 dataset and different types of features extracted: (a) original RGB frames, (b) Motion History Images, (c) forward Motion History Images, (d) backward Motion History Images, (e) SSM descriptors.

3.4.1 Datasets

We consider three different datasets:

- The *IXMAS dataset* Weinland *et al.* [2007] consists of 12 action classes (*e.g. check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point* and *pick up*). Each action is executed three times by 12 subjects and is recorded with five cameras observing the subjects from very different perspectives. The frame rate is 23 fps and the frame size 390×291 pixels.

- The *NIXMAS dataset* Weinland *et al.* [2010] contains new videos showing the same actions as in the IXMAS dataset. The dataset is recorded with different actors, cameras, and viewpoints, and about 2/3 of the videos have objects which partially occlude the actors. Overall it contains 1148 sequences.

- The *ACT*4^2 *dataset* Cheng *et al.* [2012] contains video sequences depicting 14 representative daily actions recorded through both RGB and depth channels simultaneously. The considered daily actions are: *collapse, drink, make phone call, mop floor, pick up, put on, read book, sit down, sit up, stumble, take off, throw away, twist open* and *wipe clean*.

Fig.3.3 shows some sample frames extracted from the IXMAS, NIXMAS and ACT4^2 datasets observed from different camera viewpoints. In our experiments, we use all the IXMAS and NIXMAS datasets and a subset (videos corresponding to 10 actors) of the ACT4^2 dataset.

3.4.2 Feature Representation

As discussed in Section 3.3.2, our approach is based on SSM descriptors computed from low-level features extracted on single frames. To obtain features from RGB image sequences, we used HOG and HOF features Laptev *et al.* [2008] in case of the IXMAS and NIXMAS datasets, while only HOG features are used for the $ACT4^2$ dataset. Moreover, in case of the $ACT4^2$ dataset, MHIs Bobick & Davis [2001] are used to

compute features on depth images. In an MHI, each pixel intensity is a function of the motion history at that location, where brighter value corresponds to more recent motion. Denoting by $D(x, y, t)$ the depth value corresponding to a pixel at location x, y and at time t, the MHI is computed as:

$$H_\tau^D(x,y,t) = \begin{cases} \tau, & if\ |D(x,y,t) - D(x,y,t-1)| > \delta D_{th} \\ \max(0, H_\tau^D(x,y,t-1) - 1), & otherwise \end{cases}$$

where τ is the longest time window that the system considers (τ is set equal to the number of video frames in our experiments) and δD_{th} is the threshold value for generating the mask for a motion region.

Moreover, in order to benefit to the highest degree from the depth information, two other MHI descriptors, named as forward-MHIs $H_\tau^{fD}(x, y, t)$ (encoding the information about the increase of depth) and backward-DMHIs $H_\tau^{bD}(x, y, t)$ (decrease of depth) Ni *et al.* [2011], are defined:

$$H_\tau^{fD}(x,y,t) = \begin{cases} \tau, & if\ D(x,y,t) - D(x,y,t-1) > \delta D_{th} \\ \max(0, H_\tau^{fD}(x,y,t-1) - 1), & otherwise \end{cases}$$

$$H_\tau^{bD}(x,y,t) = \begin{cases} \tau, & if\ D(x,y,t) - D(x,y,t-1) < -\delta D_{th} \\ \max(0, H_\tau^{bD}(x,y,t-1) - 1), & otherwise \end{cases}$$

To represent each video of the ACT4^2 dataset, we computed separate SSM descriptors for HOG, MHI, forward-DMHI and backward-DMHI features and, applying a bag-of words model (using 500 words), we constructed a 2000-bin histogram corresponding to the final descriptor. In Fig.3.4, the extracted MHI, forward-MHI and backward-MHI features and the corresponding SSM descriptors are shown.

3.4.3 Experimental Setup

A leave-user-out strategy was employed in our classification experiments: videos of one actor were selected for testing, while videos of the remaining actors were used as training data. For all the methods, the optimal values of the regularization parameters were determined using a separate validation set and testing the values in the interval $[2^{-6}, 2^{-5}, ..., 2^6]$. In the following, we refer to our approaches as MT-SGG-LDA

for Multi-task Sparse Graph Guided LDA, and MT-FGG-LDA for Multi-task Flexible Graph Guided LDA.

We evaluated the effectiveness of our algorithms in two cases:

- ***Multi-view Feature Sharing benefit***: Training samples from all camera views were used in this setting. According to the MTL theory, all correlated tasks are learned together. This should consequently boost each individual task's performance. Specifically, once $\mathbf{U}$ is learned for MT-SGG-LDA and $\mathbf{C}, \mathbf{S}$ are learned for MT-FGG-LDA, the test sample x_{test} is projected into C dimensional output space through the operation $x'_{test}\mathbf{u}_t$ for MT-SGG-LDA, and through $x'_{test}(\mathbf{c}_t+\mathbf{s}_t)$ for MT-SGG-LDA using $\mathbf{u}_t$ and $\mathbf{c}_t + \mathbf{s}_t$ corresponding to test view. The class label of the test sample is assigned using k-nearest neighbor classification.

- ***View-invariant Recognition benefit***: Images corresponding to one camera view were missing in the training data, and we used the model learned with images from other views to perform prediction on the missing view. In practice the test sample x_{test} is projected into a $(R-1)C$ dimensional output space through the $x'_{test}\mathbf{U}$ operation for MT-SGG-LDA and through $x'_{test}(\mathbf{C}+\mathbf{S})$ for MT-FGG-LDA since only $R-1$ views are considered in this setting. The label of the test sample is again assigned using k-nearest neighbor classification.

Table 3.1: Multi-view action recognition accuracy: comparing single and multi task learning on the IXMAS dataset.

Training with All Cameras						
	Cam1	Cam2	Cam3	Cam4	Cam5	Avg
MT-SGG-LDA	0.900	0.854	0.812	0.793	0.763	0.825
MT-FGG-LDA	0.912	0.877	0.821	0.815	0.791	0.843
Junejo - SVM Junejo *et al.* [2011]	0.748	0.745	0.748	0.706	0.612	0.727
ℓ_{12} MTL Argyriou *et al.* [2007]	0.819	0.830	0.809	0.756	0.693	0.782

3.4.4 Quantitative Evaluation

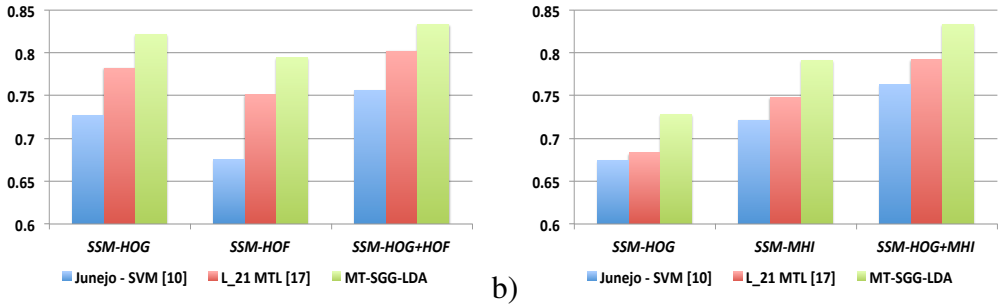
A first series of experiments were devoted to demonstrate the advantage of using an MTL approach for multi-view action recognition. To this end, we compared the pro-

Table 3.2: Multi-view action recognition accuracy: comparing single and multi task learning on the NIXMAS dataset.

Training with All Cameras						
	Cam1	Cam2	Cam3	Cam4	Cam5	Avg
MT-SGG-LDA	0.874	0.834	0.791	0.769	0.733	0.800
MT-FGG-LDA	0.888	0.841	0.799	0.785	0.764	0.815
Junejo - SVM Junejo *et al.* [2011]	0.712	0.721	0.708	0.693	0.633	0.693
ℓ_{12} MTL Argyriou *et al.* [2007]	0.803	0.794	0.768	0.763	0.672	0.760

Table 3.3: Multi-view action recognition accuracy: comparing single and multi task learning on the ACT4^2 dataset.

Training with All Cameras					
	Cam1	Cam2	Cam3	Cam4	Avg
MT-SGG-LDA	0.867	0.853	0.804	0.808	0.833
MT-FGG-LDA	0.846	0.867	0.800	0.821	0.834
Junejo - SVM Junejo *et al.* [2011]	0.799	0.787	0.743	0.721	0.763
ℓ_{12} MTL Argyriou *et al.* [2007]	0.831	0.805	0.779	0.752	0.792



a) b)

Figure 3.5: Recognition accuracy with different SSM features on the (a) IXMAS and (b) ACT4^2 datasets.

posed methods with a single SVM classifier Junejo *et al.* [2011], and the $\ell_{2,1}$-norm multi-task learning approach Argyriou *et al.* [2007] which assumes all the tasks to be related (no graph explicitly specifying task relationships is considered). In the SVM experiments, an RBF kernel was chosen and the LIBSVM[1] software package was used. A publicly available code[2] was used for $\ell_{2,1}$ multi-task learning.

Table 3.1 shows the comparison results for the IXMAS dataset. Evidently, sharing similarity information among different views using MTL is beneficial as the proposed approaches outperform SVM by at least 10%. Moreover, employing a graph modeling the degree of similarity of different views is better than assuming that the data from all views are related as in the $\ell_{2,1}$-norm MTL approach. The viewpoint associated to CAM5 is significantly different from the other four views. However, even in this case, multi-task learning is greatly beneficial as action recognition accuracy improved from 69.3% to 76-79%, implying that CAM5 view features were 'enhanced' by discriminative information from the other views. These observations show the benefit of *feature sharing* among different views achieved by our MTL framework.

Similar results were also obtained for the other two datasets as shown in Tables 3.2 and 3.3. Comparing the proposed approaches, we observe similar performance for MT-SGG-LDA and MT-FGG-LDA, with a slightly superior accuracy with the latter. We believe that this is due to the greater flexibility of the model achieved by learning the graph structure. Additionally for the IXMAS and the ACT4^2 datasets, we also evaluated the effectiveness of different features. Figure 3.5(a) shows the results on IXMAS videos obtained using various SSM descriptors computed with HOG, HOF and HOG+HOF features. As expected, the best performance was achieved using HOG + HOF features. Similarly for the ACT4^2 dataset, combining HOG+MHI proved beneficial in term of performance (Fig. 3.5(b)). This demonstrates that having at disposal not only traditional cameras but also information from the depth channel greatly improves multi-view action recognition performance. This is in accordance to what found in Cheng *et al.* [2012] where different features (extending LBP descriptors to describe depth images) are used.

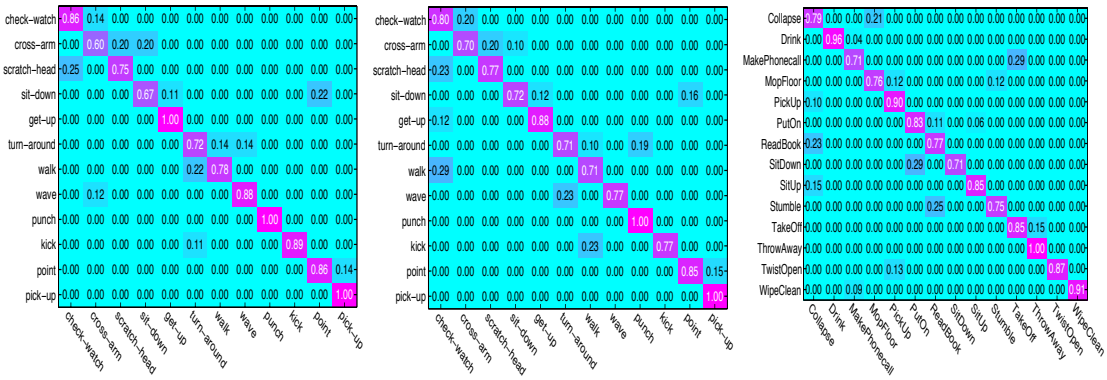
Figure 3.6 shows the confusion matrices for MT-SGG-LDA for the multi-view feature sharing experiments on the IXMAS, NIXMAS, ACT4^2 datasets respectively. By

[1] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/
[2] http://ttic.uchicago.edu/∼argyriou/code/index.html

Table 3.4: Multi-view action recognition accuracy: comparison of different methods on IXMAS dataset.

Training with All Cameras						
	Cam1	Cam2	Cam3	Cam4	Cam5	Avg
MT-SGG-LDA	0.900	0.854	0.812	0.793	0.763	0.825
MT-FGG-LDA	0.912	0.877	0.821	0.815	0.791	0.843
Li Li & Zickler [2012]	0.834	0.799	0.820	0.853	0.755	0.812
Liu Liu *et al.* [2011]	0.790	0.747	0.752	0.764	0.712	0.753
Huang Huang *et al.* [2012a]	0.632	0.586	0.604	0.568	0.476	0.573
Weinland Weinland *et al.* [2007]	0.654	0.700	0.543	0.660	0.336	0.579
Reddy Reddy *et al.* [2009]	0.696	0.692	0.620	0.651	-	0.726
Farhadi Farhadi & Tabrizi [2008]	-	-	-	-	-	0.581
Li Li *et al.* [2012]	0.910	0.919	0.911	0.906	0.871	0.905
Wu Wu & Jia [2012]	0.909	0.854	0.888	0.909	0.881	0.888
Mahasseni Mahasseni & Todorovic [2013]	0.818	0.812	0.848	0.836	0.782	0.820

observing the matrices for the IXMAS and NIXMAS datasets, it is interesting to notice that for some actions such as 'get up', 'pick up' and 'punch', our method achieves very high recognition accuracies. Even for some challenging actions (*e.g.*, 'point', 'check watch' and 'wave') having small and ambiguous motions, our method still produces reasonable and promising results.



(a) (b) (c)

Figure 3.6: Confusion matrices (MT-SGG-LDA) on the (a) IXMAS, (b) NIXMAS and (c) ACT4^2 datasets.

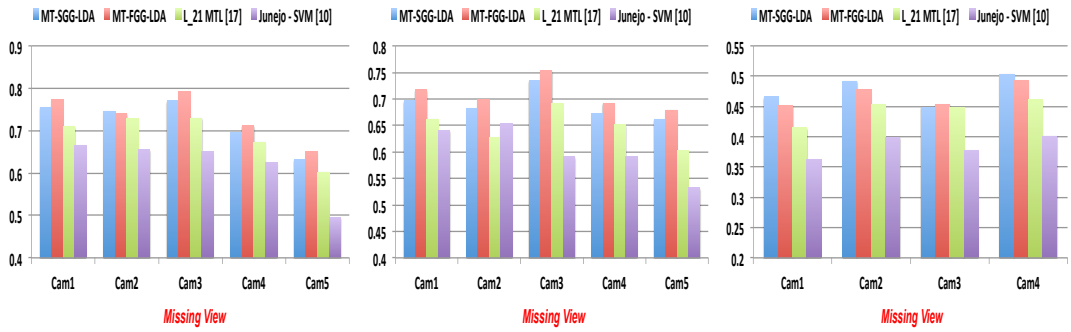
We also compared the proposed methods with other action recognition algorithms which are not based on SSMs. The results of such comparison on the IXMAS dataset

Table 3.5: Multi-view action recognition accuracy: comparison of different methods on NIXMAS dataset.

Training with All Cameras						
	Cam1	Cam2	Cam3	Cam4	Cam5	Avg
MT-SGG-LDA	0.874	0.834	0.791	0.769	0.733	0.800
MT-FGG-LDA	0.888	0.841	0.799	0.785	0.764	0.815
Weinland Weinland *et al.* [2010]	-	-	-	-	-	0.767
Mahasseni Mahasseni & Todorovic [2013]	0.782	0.816	0.807	0.776	0.761	0.788

are shown in Table 3.4. Our approach achieves higher recognition accuracy, with respect to both individual-view and (average) multi-view accuracies when compared to most previous methods. On the other hand, the approaches proposed in Li *et al.* [2012] and Wu & Jia [2012] achieve higher recognition as compared to our methods on the IXMAS dataset. However, the method in Wu & Jia [2012] is based on latent kernelized structural SVM which is intractable for inference on large-scale datasets. The feature extraction phase of the algorithm in Li *et al.* [2012] is also computationally demanding. Differently, our method is computationally efficient and easy to implement. Table 3.5 shows a similar comparison on the NIXMAS dataset. Our approaches outperform Weinland *et al.* [2010] by up to 4%, and Mahasseni & Todorovic [2013] by up to 2%. Interestingly, the latter is also based on multi-task learning. Similar results are not reported for the ACT4^2 dataset as the same setup used in the experiments in Cheng *et al.* [2012] cannot be exactly reproduced (only the videos corresponding to a subset of actors are used for evaluation in Cheng *et al.* [2012]).

Finally, to demonstrate the benefits of our approach on *view-invariant* action recognition, we evaluated its performance when one view was missing in the training data. Results on the IXMAS, NIXMAS and ACT4^2 datasets are shown in Fig.3.7(a), (b) and (c) respectively. Although there is some performance drop compared to the situation where all camera views are available in the training phase, our approach still achieves better performance than the single-task SVM and $\ell_{2,1}$ multi-task learning methods. The recognition accuracy of both our approaches are similar, with MT-SGG-LDA outperforming MT-FGG-LDA in the experiments on the ACT4^2 dataset. This may be due to the importance of sparsity when the size of the feature vectors increases.



(a) (b) (c)

Figure 3.7: Cross-view action recognition accuracy: training is performed with one view missing on (a) IXMAS, (b) NIXMAS, (c) ACT4^2 datasets.

3.5 Conclusions

We have considered the problem of multi-view human action recognition in this work and proposed a multi-task extension of multi-class LDA to effectively address the same. Experimental results on the IXMAS, NIXMAS and ACT4^2 datasets demonstrate the superior performance of our method compared to other SSM-based state-of-the-art methods. Possible future works include the integration of other view-invariant features in combination with SSM descriptors and the investigation of a different strategy for graph construction based on camera geometry information.

Chapter 4

Coupling GLocal Structural for Feature Selection with Sparsity for Image and Video Classification

The selection of discriminative features is an important and effective technique for many computer vision and multimedia tasks. Using irrelevant features in classification or clustering tasks could deteriorate the performance. Thus, designing efficient feature selection algorithms to remove the irrelevant features is a possible way to improve the classification or clustering performance. With the successful usage of sparse models in image and video classification and understanding, imposing structural sparsity in *feature selection* has been widely investigated during the past years. Motivated by the merit of sparse models, in this paper we propose a novel feature selection method using a sparse model. Different from the state of the art, our method is built upon $\ell_{2,p}$-norm and simultaneously considers both the global and local (GLocal) structures of data distribution. Our method is more flexible in selecting the discriminating features as it is able to control the degree of sparseness. Moreover, considering both global and local structures of data distribution makes our feature selection process more effective. An efficient algorithm is proposed to solve the $\ell_{2,p}$-norm joint sparsity optimization problem in this paper. Experimental results performed on real-world image and video datasets show the effectiveness of our feature selection method compared to several state-of-the-art methods.

4.1 Introduction

Many applications in computer vision and multimedia, such as image and video annotation, require images and videos be represented by low-level features. If some of these features are irrelevant or redundant, this could be deleterious for the performance of classification or clustering tasks. The main idea of feature selection is to choose a subset of input variables by eliminating the features with little or no predictive information. By removing such features from the original feature representation, feature selection could speed up the learning process, enhance model generalization capability and alleviate the effect of curse of dimensionality.

In the past several years, there have been many feature selection methods proposed in the computer vision, pattern recognition and multimedia communities. Typically, feature selection algorithms fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric and eliminates the features that do not achieve an adequate score. Subset selection searches the set of possible features for the optimal subset. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right. For example, a physician may make a decision based on the selected features whether a dangerous surgery is necessary for treatment or not.

In computer vision and multimedia areas, feature selection based on subset selection has drawn more attention. The classical Fisher Score Duda *et al.* [2001] feature selection method evaluates the relevance of features according to the label distribution of the data points. Minimum-Redundancy-Max-Relevance Peng *et al.* [2005] feature selection method selects useful features which have the strongest correlation with a classification variable based on mutual information. Several previous works have shown the effectiveness of these methods. However, these traditional algorithms usually evaluate features one by one, which is not computationally efficient and ignores the correlation between different features.

Recently, sparse models have been successfully used in the multimedia and computer vision tasks such as image classification Moxley *et al.* [2010]; Wang *et al.* [2009]; Yuan & Yan [2010], headpose estimation Yan *et al.* [2013b], face recognition Wagner

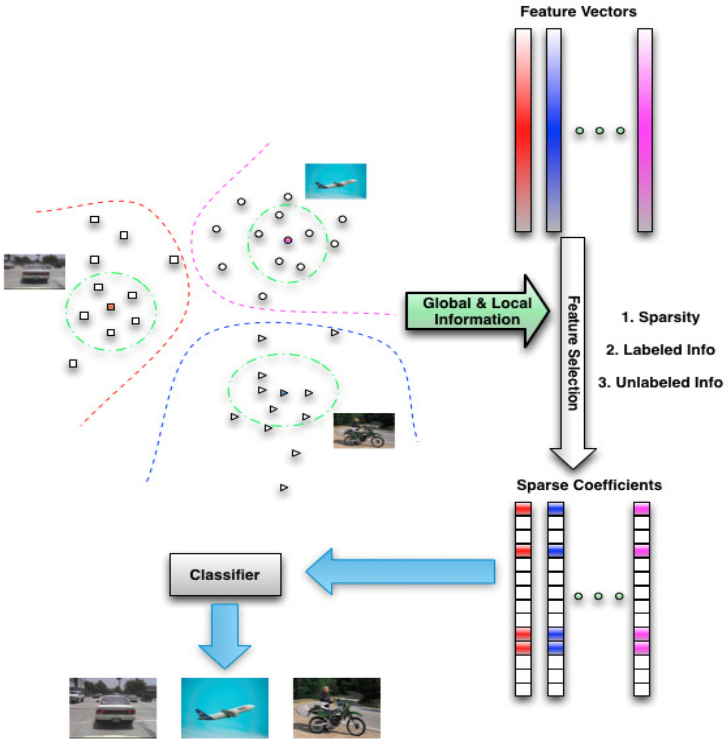


Figure 4.1: Our GLocal Structural feature selection with Sparsity (GLSS) Framework

et al. [2012]; Wright *et al.* [2009] and action recognition Yan *et al.* [2013a]. The sparse model has also been widely investigated in feature selection. The intuition for this type of approaches is that many real-world data are often sparse which means feature selection can be achieved by searching the sparse representation of the data. In contrast with the traditional feature selection approach by using the sparse model, we are able to select features jointly in a batch mode and meanwhile to leverage the correlation between different features. Nie *et al.* Nie *et al.* [2010] leverage joint $\ell_{2,1}$-norm minimization on both loss function and regularization for feature selection. Yang *et al.* Yang *et al.* [2011] have proposed an $\ell_{2,1}$-norm regularized feature selection method for unsupervised learning considering the manifold structure of data representation. Yang *et al.* Yang *et al.* [2012] proposed a semi-supervised algorithm called ranking with Local Regression and Global Alignment (LRGA) to learn a robust Laplacian matrix for data ranking. Gao *et al.* Gao *et al.* [2013] proposed a method which simultaneously utilized both visual and textual information to estimate the relevance of user tagged images.

The sparse model for feature selection has proved to be effective but most existing methods have two limitations. On one hand, the widely used ℓ_1-norm and $\ell_{2,1}$-norm are not flexible enough to control the sparseness of feature representation. Hence, useful features could be neglected or noisy features could be selected. In other words, they may not find out the optimal subset of the original features. On the other hand, many of them only consider global information of the data representation but ignore the local structure of data distribution which also has significantly useful information for selecting more discriminating features. Locality Preserving Projections (LPP) He & Niyogi [2003] searches for an embedding space in which the similarity among the local neighborhoods is preserved. However, LPP has two disadvantages: Firstly, LPP does not take the label information into consideration which is crucial for classification tasks; Secondly, like most graph-based methods, graph construction of LPP is sensitive to noise and outliers. To address these issues, in this paper we propose a novel and robust feature selection method by employing a $\ell_{2,p}$-norm based sparse model and meanwhile considering both global and local data structures. We name our method GLocal Structural feature selection with Sparsity (GLSS). Instead of using the traditional ℓ_1-norm or $\ell_{2,1}$-norm, we propose to exploit the $\ell_{2,p}$-norm based sparse model. Since we can adjust the value of p in our framework, our algorithm is more flexible to

control the sparseness of the feature representation, thus resulting in a better subset of the original feature set. To use the information of both global and local data structure, we build two regression models in a joint framework: one for all the data points and one for the local neighboring data points. Figure 1 illustrates the overview of our feature selection method. All training and test samples are represented by low-level feature vectors. Our GLSS method considers global and local information with sparsity. Then the selected features are fed into a classifier to do image or video classification.

The main contributions of our work are as follows:

- Our method GLSS utilizes the $\ell_{2,p}$-norm based sparse model for feature selection. This model is more capable of selecting discriminative features by adjusting the value of p;
- GLSS is built upon both global and local data structures. Exploring the GLocal information helps boosting the efficacy of feature selection.

This paper is the extension of our conference paper Yan *et al.* [2013c]. The extension includes both the theoretical principle and the applications. We would like to highlight them as follow:

- We add more details about the intuitive of our objective function and the formulation derivative.
- We used our method for two more applications, which are video concept detection and image annotation. Our method shows promising performance and is especially competitive when few labeled samples are available, which makes it attractive for large scale multimedia data understanding.
- We conducted more complementary analyzing experiments on the 6 datasets to assess the overall performance of our method for different applications. These include studies which aim to understand the influence of the sparse level, the influence of the unlabeled data, the influence of the local set and parameter sensitivity studies which demonstrates how the parameters affect the performance.
- We compared our method with another $L_{2,1}$-norm feature selection method. The experiment results show great improvement over $L_{2,1}$-norm feature selec-

tion method which prove the advantage of our $L_{2,p}$-norm sparsity level adjusted method.

- For practical applications it is interesting how fast our algorithm converges. Therefore, we also conducted a newly added experiment which studies the convergence of our method.

The rest of paper is organized as follows. In section II, we illustrate the formulation of our framework and propose an algorithm for solving the objective function. Experiments are given in section III and section IV draws the conclusion of this paper.

4.2 GLocal Structural Feature Selection with Sparsity

The low-level features of images or videos incorporate different information, either globally or locally. Intuitively, effective analysis on both levels would boost the feature selection efficacy. In this section, we propose our GLocal Structural feature selection with Sparsity (GLSS) algorithm and derive an effcient solver for the problem.

4.2.1 Problem Formulation

We first explore the local data structure to help the selection of discriminating features from the original representation. Inspired by previous works Yang *et al.* [2009], Zhang & Zha [2002], we build a local set $\mathcal{N}_i$ for each datum x_i. $\mathcal{N}_i = \{x_i, x_{i1}, ..., x_{ik-1}\}$ and it consists of x_i and its $k-1$ nearest neighbors. For each local set, a local prediction function f_i is defined to correlate the data within the set with their predicted labels and we can obtain f_i through the following objective function:

$$\sum_{x_k \in \mathcal{N}_i} \ell\left(f_i(x_k), q_k\right) + \alpha\Omega(f_i) \tag{4.1}$$

where $\ell(\cdot)$ is the loss function and $\Omega(\cdot)$ is the regularizer. $x_k \in \mathcal{N}_i$ and q_k is the predicted label for x_k. α is a regularization parameter.

Globally, we also define a prediction function f to correlate all the n data points

with their predicted labels as follows:

$$\sum_{i=1}^{n} \ell\left(f(x_i), q_i\right) + \gamma\Omega(f) \tag{4.2}$$

where q_i is the predicted label for x_i and γ is a regularization parameter.

Suppose the feature dimension is d and there are c classes. We apply linear regression model and obtain $f_i(x) = W_i^T x + b_i$ and $f(x) = W^T x + b$ where $W_i \in \mathbb{R}^{d\times c}$ and $W \in \mathbb{R}^{d\times c}$ are two projection matrices. We aim to leverage both the global and local information Yan *et al.* [2013c]. Hence, we propose the following joint framework:

$$\begin{aligned}\min_{W,W_i,q_i,q_j,b,b_i} \sum_{i=1}^{n}\sum_{x_j\in\mathcal{N}_i} (\left\|W_i^T x_j + b_i - q_j\right\|_F^2 + \alpha\left\|W_i\right\|_F^2) \\ +\beta(\sum_{i=1}^{n}\left\|W^T x_i + b - q_i\right\|_F^2 + \alpha\Omega(W))\end{aligned} \tag{4.3}$$

where β is a parameter and $\|\cdot\|_F$ is the Frobenius norm of matrix. As W is used for feature selection, a sophisticated regularizer is needed to make W able to reflect the importance of different features. Previous work has shown that sparse models are useful for feature selection by eliminating redundancy and noise Ma *et al.* [2012], Argyriou *et al.* [2007], Obozinski *et al.* [2007]. The sparse models are used to make some of the feature coefficients shrink to zeros. As a result, W can be regarded as the combination coefficients for the most discriminative features to achieve feature selection.

Specifically, we propose to minimize $\|W\|_{2,p} = \left(\sum_{i=1}^{d}(\sum_{j=1}^{c}|W_{ij}|)^{\frac{p}{2}}\right)^{\frac{1}{p}}$ to achieve that goal. $\|\cdot\|_{2,p}$ denotes the $\ell_{2,p}$-norm ($0 < p < 2$). p is used to control the degree of sparseness. The lower p is, the more sparse W is or in other words, more rows of W would shrink to zeros. A merit of using $\ell_{2,p}$-norm, compared to using $\ell_{2,1}$-norm or ℓ_1-norm, is that we can adjust the value of p to search for the optimal subset from the

original features. Consequently, Eqn. (4.3) can be rewritten as:

$$\min_{W,W_i,q_i,q_j,b,b_i} \sum_{i=1}^{n} \sum_{x_j \in \mathcal{N}_i} (\left\| W_i^T x_j + b_i - q_j \right\|_F^2 + \alpha \left\| W_i \right\|_F^2) \\ +\beta(\sum_{i=1}^{n} \left\| W^T x_i + b - q_i \right\|_F^2 + \alpha \left\| W \right\|_{2,p}^p) \tag{4.4}$$

Let $X_i = [x_i, x_{i1}, ..., x_{ik-1}] \in \mathbb{R}^{d \times k}$, $Q_i = [q_i, q_{i1}, ..., q_{ik-1}]^T \in \mathbb{R}^{k \times c}$, Eqn. (4.4) can be rewritten as:

$$\min_{W,W_i,Q,Q_i,b,b_i} \sum_{i=1}^{n} (\left\| X_i^T W_i + 1_k b_i^T - Q_i \right\|_F^2 + \alpha \left\| W_i \right\|_F^2) \\ +\beta(\left\| X^T W + 1_n b^T - Q \right\|_F^2 + \alpha \left\| W \right\|_{2,p}^p) \tag{4.5}$$

where $1_k \in \mathbb{R}^k$ and $1_n \in \mathbb{R}^n$ are two vectors with all ones. Next, we build up the connection between the predicted labels $Q \in \mathbb{R}^{n \times c}$ and the ground truth labels $Y \in \mathbb{R}^{n \times c}$. Q is supposed to be consistent with Y and we propose to minimize $Tr\left((Q-Y)^T U (Q-Y)\right)$ inspired by Saberian *et al.* [2011]; Schlkopf *et al.* [1998]; Yang *et al.* [2007]; Zhu *et al.* [2003]. U is a diagonal matrix whose diagonal element $U_{ii} = \infty$ if x_i is labeled and $U_{ii} = 1$ otherwise. To this end, we propose the following objective function for feature selection based on both local and global data structure:

$$\min_{W,W_i,Q,Q_i,b,b_i} \sum_{i=1}^{n} (\left\| X_i^T W_i + 1_k b_i^T - Q_i \right\|_F^2 + \alpha \left\| W_i \right\|_F^2) \\ +\beta(\left\| X^T W + 1_n b^T - Q \right\|_F^2 + \alpha \left\| W \right\|_{2,p}^p) \\ +Tr\left((Q-Y)^T U (Q-Y)\right) \tag{4.6}$$

After W is learned, we can see that many rows of the optimal W shrink to zeros (close to zeros). We rank each feature according to $\left\| W \right\|_F$ in descending order and select the top ranked features.

4.2.2 Optimization

In this subsection, we present our solution for Eqn. (4.6). Since it involves the $\ell_{2,p}$-norm which is non-smooth and cannot be solved in a closed form, we adopt the alternating minimization algorithm to optimize the objective function with respect to b_i, W_i, b, Q_i, Q and W respectively in five steps as follows:

Step 1, Fix W_i**,** b**,** Q_i**,** Q**,** W**, Optimize** b_i

By setting the derivative of Eqn. (4.6) *w.r.t.* b_i to zero, we have:

$$b_i = \frac{1}{k}(Q_i^T 1_k - W_i^T X_i 1_k)$$

Step 2, Fix b_i**,** b**,** Q_i**,** Q**,** W**, Optimize** W_i

By setting the derivative of Eqn. (4.6) *w.r.t.* W_i to zero, we have:

$$W_i = (X_i H_k X_i^T + \alpha I)^{-1} X_i H_k Q_i$$

where $I \in \mathbb{R}^{d \times d}$ is an identity matrix and $H_k = I - \frac{1}{k} 1_k 1_k^T$ is a locally centering matrix.

Step 3, Fix b_i**,** W_i**,** Q_i**,** Q**,** W**, Optimize** b

By setting derivative of Eqn. (4.6) *w.r.t.* b to zero, we have:

$$b = \frac{1}{n}(Q^T 1_n - W^T X 1_n)$$

Step 4, Fix b_i**,** W_i**,** b**,** W**, Optimize** Q_i **and** Q

Denote $G_i = H_k - H_k X_i^T (X_i H_k X_i^T + \alpha I)^{-1} X_i H_k$ and define a selection matrix $S \in \{0,1\}^{n \times k}$, where $S^{ij} = 1$ if x_i is the j^{th} element in $\mathcal{N}_i$ and $S^{ij} = 0$, otherwise. Consequently, $Q_i = S^T Q$ and we can obtain:

$$\sum_{i=1}^{n} Tr(Q_i^T G_i Q_i) = Tr\left(Q^T (\sum_{i=1}^{n} S G_i S^T) Q\right) = Tr(Q^T L_l Q)$$

where $L_l = \sum_{i=1}^{n} S G_i S^T$ reflects the exploration of local data structure. In this way, Q_i

absorbed by Q which means we only need to optimize Q. Then Eqn. (4.6) becomes:

$$\min_{Q} Tr(Q^T L_l Q) + \beta \left\| X^T W + 1_n b^T - Q \right\|_F^2 \\ + Tr\left((Q-Y)^T U (Q-Y)\right)$$

By setting the derivative *w.r.t.* Q to zero, we have:

$$Q = (L_l + \beta H + U)^{-1}(UY + \beta H X^T W).$$

where $H = I - \frac{1}{n} 1_n 1_n^T$ is the global centering matrix.

Step 5, Fix b_i**,** W_i**,** Q_i**,** Q**,** b**, Optimize** W

Denoting $W = [w^1, ..., w^d]^T$, we define a diagonal matrix D with its diagonal elements $D^{ii} = \frac{1}{\frac{2}{p}\|w^i\|_2^{2-p}}$. The objective is then equivalent to:

$$\min_{W} \left\| X^T W + 1_n b^T - Q \right\|_F^2 + \alpha Tr(W^T D W)$$

By setting the derivative *w.r.t.* W to zero, we obtain:

$$W = (A + \alpha\beta D)^{-1} B$$

where $A = XH\left(\beta I - \beta^2 (L_l + U + \beta H)^{-1}\right) HX^T$ and $B = \beta XH(L_l + U + \beta H)^{-1} UY$.

According to the optimization, we propose Algorithm 4 to solve the objective function of Eqn. (4.6). The detailed convergence analysis of the algorithm is provided in Appendix.

4.3 Experiments

In this section, we conduct extensive experiments to evaluate the performance of our feature selection method. We also compare our method with other state-of-the-art feature selection methods.

Algorithm 4 The algorithm for GLocal Structural feature selection with Sparsity (GLSS).

Input: The training data $X \in \mathbb{R}^{d\times n}$; The training data labels $Y \in \mathbb{R}^{n\times c}$; Parameters α, β and p.
Output: Optimized $W \in \mathbb{R}^{d\times c}$.

Construct X_i for each datum;
Compute U;
Compute H_k, H;
Compute the selection matrix S;
Let $G_i = H_k - H_k X_i^T (X_i H_k X_i^T + \alpha I)^{-1} X_i H_k$;
Construct L_l as $L_l = \sum_{i=1}^{n} S G_i S^T$;
Set $t = 0$, initialize $W_t \in \mathbb{R}^{d\times c}$ randomly;
Compute $A = XH\left(\beta I - \beta^2 (L_l + U + \beta H)^{-1}\right) HX^T$;
Compute $B = \beta XH(L_l + U + \beta H)^{-1} UY$;
Repeat
Compute $W_t = [w_t^1, ..., w_t^d]^T$;

Compute the diagonal matrix D_t as: $D_t = \begin{bmatrix} \frac{1}{\frac{2}{p}\|w^1\|_2^{2-p}} & & \\ & \cdots & \\ & & \frac{1}{\frac{2}{p}\|w^d\|_2^{2-p}} \end{bmatrix}$;

Update W_{t+1} as: $W_{t+1} = (A + \alpha\beta D_t)^{-1} B$;

$t = t + 1$.
Until Convergence
Return W.

4.3.1 Datasets and Low Level Feature Extraction

We give a brief description of all datasets and their low level features that we used in the experiments:

- *Columbia University Image Library* (**COIL-20**) Nene *et al.* [1996]: COIL-20 dataset contains 1440 images of 20 different objects. The objects were placed on a motorized turnable. The turnable was rotated through 360 degrees to vary object pose with respect to a fixed camera. We extract 81 dimensions histogram of oriented gradients (HOG) as features.

- *COREL 50 Category Color Photos Dataset* (**COREL-50**) Hoi *et al.* [2008]: COREL-50 contains 5000 images from 50 categories (100 images per category). The images are complex color scenes taken from a large commercially available CD-ROM library allowing access to several thousand stimuli. We extract 9 dimensional color histogram, 18 dimensional edge direction histogram and 9 dimensional wavelet to concatenate into a 36 dimensional vector as low-level feature.

- *HumanEva dataset* (**HumanEva**) Sigal *et al.* [2010]: HumanEva 3D motion database contains five types of actions, namely boxing, gesturing, jogging, walking and throw-catch performed by different subjects. We randomly sample 10,000 data of two subjects (5,000 per subject). The action of the two subjects is considered to be different. We simultaneously recognize the identities and actions, which comes to 10 semantic categories in total. Each action is encoded as a collection of 16 joint coordinates in 3D space, thus resulting in a 48 dimensional feature vector. On top of that, we compute the Joint Relative Features between different joints and get a feature vector of 120 dimensions. The two kinds of feature vectors are further combined to generate a 168 dimensional feature.

- *MIR FLICKR dataset* (**MIR FLICKR**)[1] Huiskes & Lew [2008]: This image collection consists of 25000 images that were downloaded from the social photography site Flickr.com through its public API. The color images are representative of a generic domain and are of high quality. The average number of tags per image is 8.94. In the collection there are 1386 tags which occur in at least 20 images. We extract 512 dimensional GIST feature as low level representations.

- *Youtube Action dataset* (**Youtube**) Liu *et al.* [2009]: Youtube Action dataset contains 11 action categories: *basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog*. It contains large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc. We extract the STIP features and then generate a 1000 dimension Bag-of-Words feature to represent

[1] Multi-label dataset

each video sequence.

- *Kodak Consumer Video dataset* (**Kodak**)[1] Yanagawa *et al.* [2008]: Kodak Consumer Video dataset consists of 1,358 consumer video clips and 5,166 key-frames are extracted accordingly. Among these key-frames, 3590 ones are annotated. We use all the annotated key-frames belonging to 22 concepts in our experiments for video concept detection. 144 dimensional Color Correlogram are used to represent the key-frames.

The detailed description of the datasets is shown in Table 1.

Table 4.1: Dataset Description

Dataset	Size	# of Features	# of Classes
Youtube Liu *et al.* [2009]	1596	1000	11
Kodak Yanagawa *et al.* [2008]	3590	144	22
HumanEva Sigal *et al.* [2010]	10000	168	10
MIR FLICKR Huiskes & Lew [2008]	25000	512	38
COIL-20 Nene *et al.* [1996]	1440	81	20
COREL-50 Hoi *et al.* [2008]	5000	36	50

4.3.2 Comparison Methods

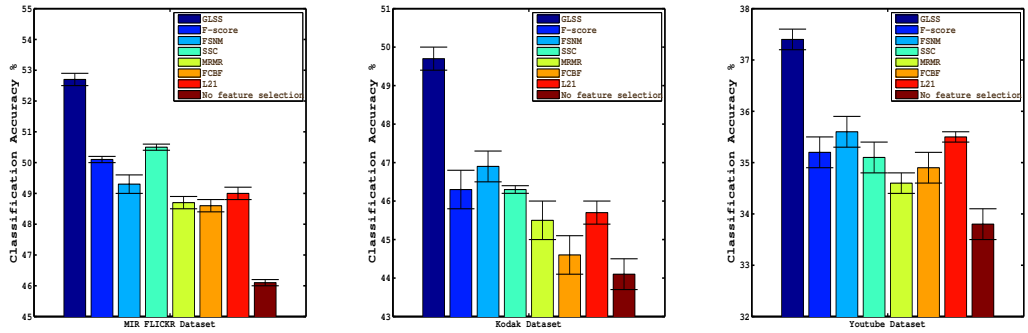
We compare our GLSS feature selection method with several feature selection algorithms for image and video classification described as follows:

- *Baseline* (**No Feature Selection**): Classification without using any feature selection method. This is used as the baseline for all the experiments.

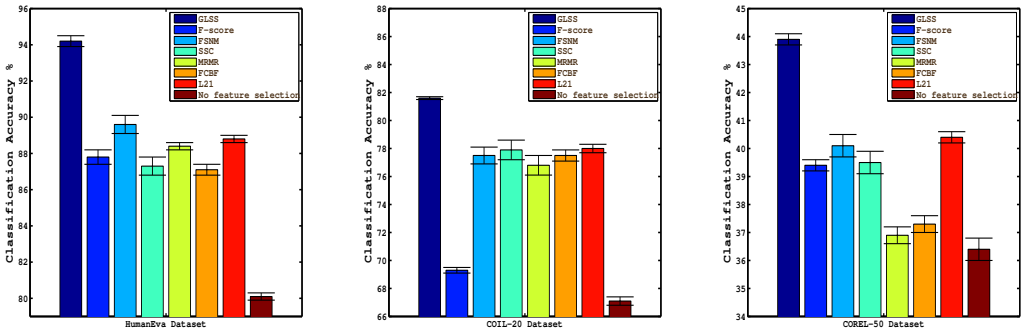
- *Fisher Score* (**F-score**) Duda *et al.* [2001]: A classical feature selection algorithm which is widely used in literature.

- *Feature selection with Joint $\ell_{2,1}$-norm minimization* (**FSNM**) Nie *et al.* [2010]: It employs joint $\ell_{2,1}$-norm minimization on both loss function and regularization to realize feature selection across all data points.

- *Sparse Multinomial Logistic Regression via Bayesian L1 Regularization* (**SBMLR**) Cawley *et al.* [2006]: It exploits sparsity by using a Laplace prior and is used for multi-class pattern recognition. It can also be utilized for feature selection.



(a) (b) (c)



(d) (e) (f)

Figure 4.2: Classification Accuracy Comparision (MAP±Standard Deviation) when 10% training samples are labeled on (a) MIR FLICKR dataset (b) Kodak dataset (c) Youtube Action dataset (d) HumanEva dataset (e) COIL-20 dataset (f) COREL-50 dataset.

- *Minimum-Redundancy-Maximum-Relevance* (**MRMR**) Peng *et al.* [2005]: It uses either mutual information, correlation, distance or similarity scores to select features. Relevant features and redundant features are considered simultaneously with mutual information.

- *Fast Correlation-Based Filter* (**FCBF**) Liu & Yu [2003]: This is an algorithm designed for high-dimensional data and has been shown effective in removing both irrelevant and redundant features.

- $\ell_{2,1}$-*norm Manifold* (**L21**) Yang *et al.* [2011]: This is an $\ell_{2,1}$-norm regularized feature selection method for unsupervised learning considering the manifold structure of data representation.

4.3.3 Experimental Setup

In our experiments, each feature selection algorithm is first used to select features. Then the KNN classifier (K=5) is applied based on the selected features. Multi-label KNN Zhang & Zhou [2007] is adopted when it is a multi-label classification problem (i.e., for MIR Flicker and Kodak). The original dataset is randomly partitioned into 3 subsets. Of the 3 subsets, a single subset is retained as the test data and the remaining 2 subsets are used as training data. We use 3-fold cross validation in all the experiments and report the average results with standard deviation.

The parameters α and β in Eqn. (4.6) are tuned from $\{0.01, 1, 100\}$. The parameter p is tuned form $\{0.05, 0.5, 1\}$. We tune the regularization parameters for other compared algorithms similarly. For all the algorithms in our experiments, the number of selected features is set as $\{400, 600, 800\}$ for Youtube Action dataset, $\{40, 80, 120\}$ for Kodak and HumanEva dataset, $\{150, 250, 350\}$ for the MIR FLICKR dataset, $\{25, 50, 75\}$ for the COIL-20 dataset, and $\{10, 20, 30\}$ for the COREL-50 dataset. We report the best results of all the algorithms using different parameters.

4.3.4 Comparision with other methods

Figure 4.2 shows the classification results using different methods based on different datasets when 10% of the training samples are labeled. Our GLSS outperforms

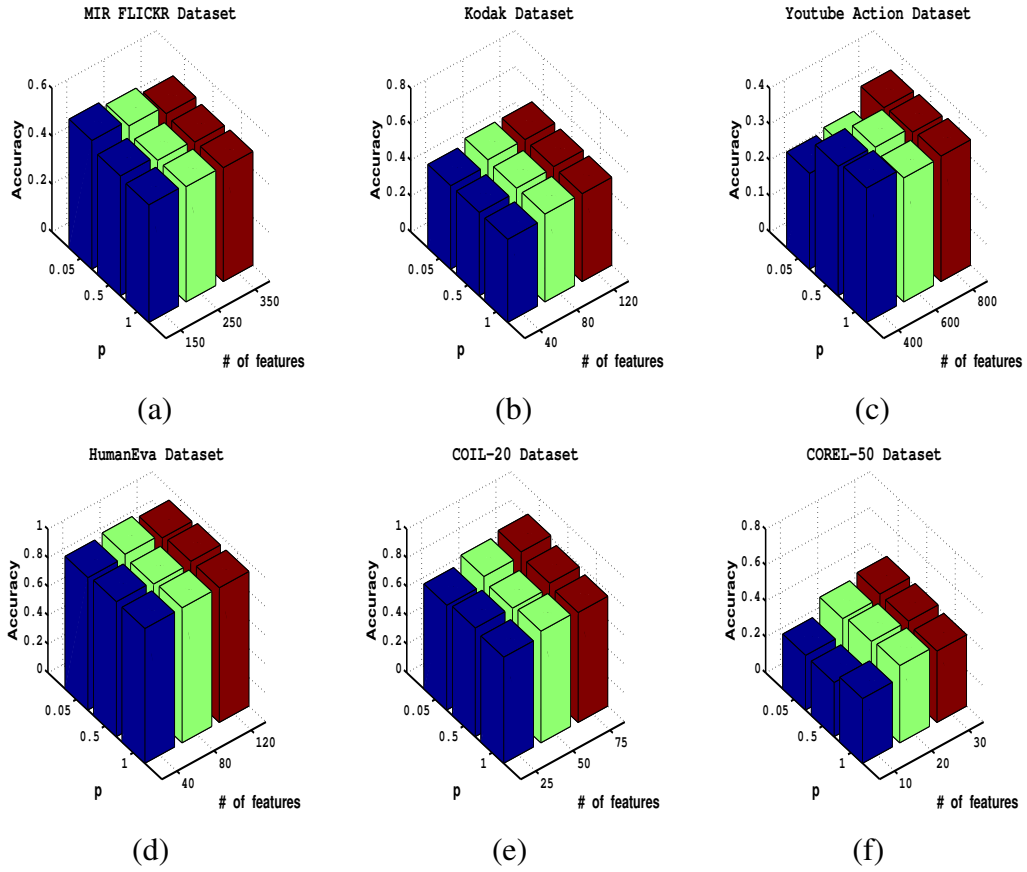


Figure 4.3: Performance variance $w.r.t\ p$ and $\#\ of\ features$ on (a) MIR FLICKR dataset (b) Kodak dataset (c) Youtube Action dataset (d) HumanEva dataset (e) COIL-20 dataset (f) COREL-50 dataset.

other methods on all six datasets. 3%, 3%, 2%, 5%, 4%, 4% more accuracy has been achieved on MIR FLICKR, Kodak, Youtube Action, HumanEva, COIL-20 and COREL-50 datasets respectively over the second best feature selection method. 6%, 6%, 3%, 14%, 13%, 8% more accuracy has been achieved respectively when compared to the baseline (i.e., not using feature selection). The results demonstrate the advantage of our GLSS method. GLSS performs better than $\ell_{2,1}$-norm Yang *et al.* [2011] since GLSS could control the feature sparsity level through $\ell_{2,p}$-norm. Moveover, GLSS considers both global and local structure of data distribution which are both critical in classification tasks. Therefore, GLSS has shown better performance than all other feature selection methods Nie *et al.* [2010], Duda *et al.* [2001] which do not consider local structure of data distribution. Our method benefits from these two important factors which gives our method the best performance on six different datasets.

4.3.5 Classifiers Effect Analysis

To understand the KNN classifier effect for our proposed GLSS feature selection method, we evaluate different values for the neighborhood parameters N for the KNN classifiers (10% of the training samples are labeled). Table 2 shows the classification accuracy (GLSS method accuracy/No feature selection accuracy) on several datasets. Multi-label classifier is adopted when it is necessary. We can see from Table 2 that the largest performance gain is acchieved when N = 5 compared with simply classifying without any feature selection method tested on four different datasets.

Table 4.2: KNN Classifier Effect

GLSS/No-FS	N = 3	N = 5	N = 7
COREL-50 Hoi *et al.* [2008]	43.1%/36.3%	43.9%/36.7%	43.7%/36.0%
Kodak Yanagawa *et al.* [2008]	47.7%/44.7%	49.3%/45.7%	47.5%/44.3%
HumanEva Sigal *et al.* [2010]	92.7%/79.8%	94.2%/80.1%	93.2%/79.1%
MIR FLICKR Huiskes & Lew [2008]	51.3%/45.1%	52.7%/46.1%	51.8%/45.0%

We also use different classifiers to evaluate our proposed GLSS feature selection method (10% of the training samples are labeled). Table 3 shows the classification accuracy (GLSS method accuracy/No feature selection accuracy) on several datasets based on three different kinds of classifiers, e.g. KNN, SVM and Adaboost. Multi-label classifier is adopted when it is necessary. We can see from Table 3 that SVM classifier is less sensitive to feature selection. However, our GLSS still achieves at

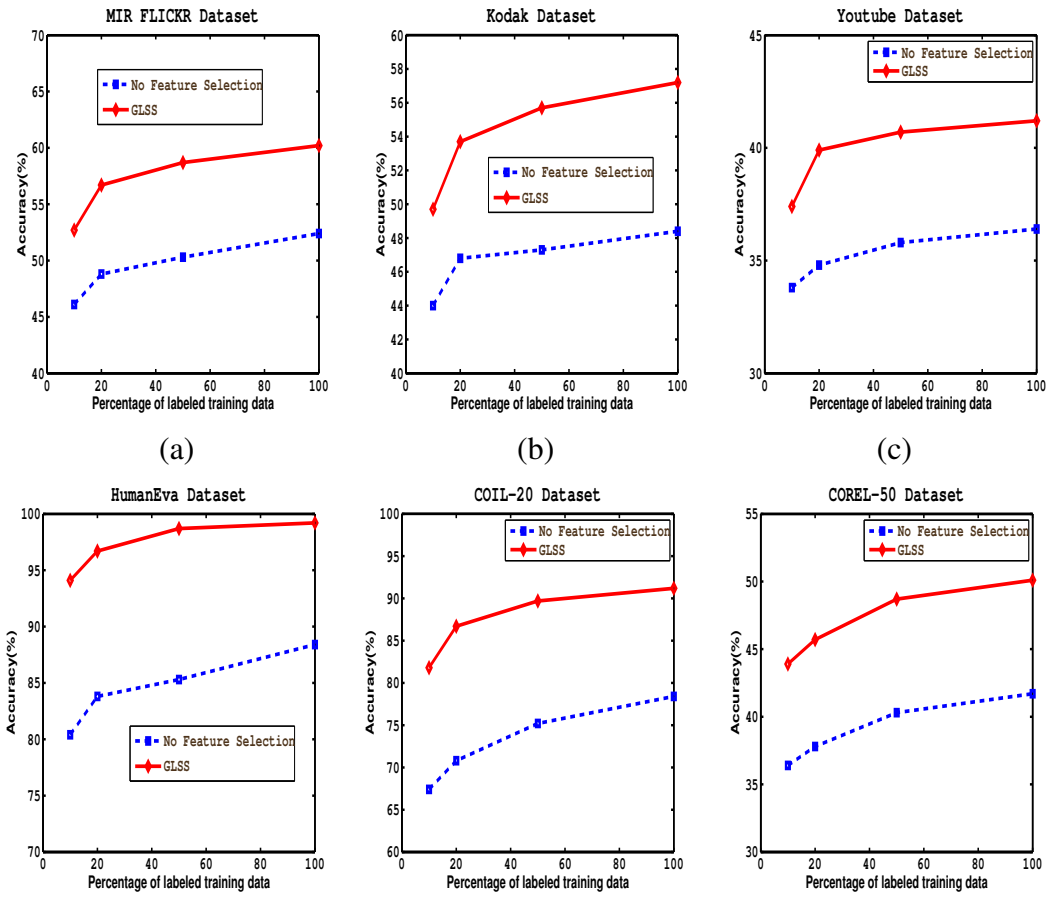


Figure 4.4: Performance variance $w.r.t.$ the percentage of labeled training data on (a) MIR FLICKR dataset (b) Kodak dataset (c) Youtube Action dataset (d) HumanEva dataset (e) COIL-20 dataset (f) COREL-50 dataset.

least 4% accuracy performance gain compared with simply classifying without any feature selection method tested on 6 different datasets.

Table 4.3: Different Type of Classifiers Effect

GLSS/No-FS	KNN[2]	SVM[3]	Adaboost[4]
Youtube Liu *et al.* [2009]	37.4%/33.8%	38.2%/34.3%	37.6%/33.7%
Kodak Yanagawa *et al.* [2008]	49.3%/45.7%	49.3%/45.9%	49.5%/45.3%
HumanEva Sigal *et al.* [2010]	94.2%/80.1%	95.7%/83.4%	94.5%/81.0%
MIR FLICKR Huiskes & Lew [2008]	52.7%/46.1%	53.0%/46.5%	52.6%/46.0%
COIL-20 Nene *et al.* [1996]	81.6%/67.1%	82.2%/68.7%	81.9%/68.0%
COREL-50 Hoi *et al.* [2008]	43.9%/36.7%	44.3%/38.2%	43.7%/36.9%

4.3.6 Sparsity Analysis

To understand the influence of parameter p and the number of features selected in our method, we perform an experiment on the parameter sensitivity. Figure 4.3 shows the classification accuracy $w.r.t.$ p and the number of features selected when we fix parameters α and β. We can see that the performance is more sensitive to sparsity control parameter p compared to the number of features selected in general. However, the performance is more sensitive to the number of features selected when parameter p is small (0.05 in our experiments). The best performance is always achieved when we obtain the most sparsest situation (p = 0.05) in our experiments. Table 4.4 illustrates the sparsity level of W (measured by the number of rows of W shrinking to zeros) based on different p-norm. We observe that a smaller p induces a sparser W.

Table 4.4: Sparsity level of W based on different p-norm. (# rows of W that shrink to zeros)

	p = 0.05	p = 0.5	p = 1
MIR FLICKR Huiskes & Lew [2008]	105	77	51
Youtube Liu *et al.* [2009]	454	311	105
Kodak Yanagawa *et al.* [2008]	77	53	36

[2]http://lamda.nju.edu.cn/code_MLkNN.ashx
[3]http://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/multilabel/
[4]http://cse.seu.edu.cn/people/zhangml/Resources.htm/

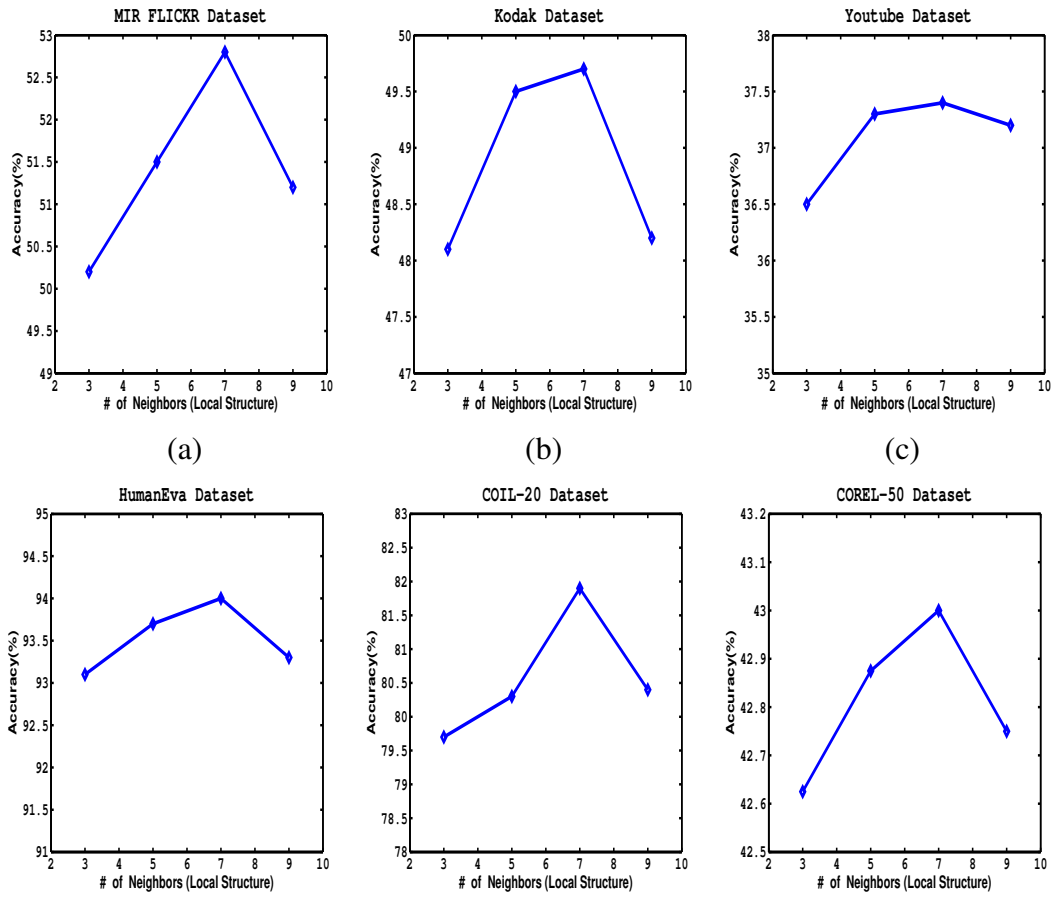


Figure 4.5: Performance variance $w.r.t.$ the number of Local Structure Neighbors on (a) MIR FLICKR dataset (b) Kodak dataset (c) Youtube Action dataset (d) HumanEva dataset (e) COIL-20 dataset (f) COREL-50 dataset.

4.3.7 Global and Local Effect

To investigate the contribution of local structure to the proposed problem, we conduct experiments with and without the first term in Eqn. (4.6). The experimental results are listed in Table 4.5. From Table 4.5, we observe that considering both global and local data structure could improve the performance at least 5% on six different datasets comparing with considering solo global data structure. This is intuitive since the local structure of data distribution also has significantly useful information for selecting more discriminating features.

Table 4.5: Global and Local Effect (SVM classifier)

	Global	Global & Local
Youtube Liu *et al.* [2009]	34.3%	38.2%
Kodak Yanagawa *et al.* [2008]	44.9%	49.3%
HumanEva Sigal *et al.* [2010]	90.1%	95.7%
MIR FLICKR Huiskes & Lew [2008]	48.7%	53.0%
COIL-20 Nene *et al.* [1996]	77.6%	82.2%
COREL-50 Hoi *et al.* [2008]	39.8%	44.3%

4.3.8 Influence of the Unlabeled Data

To evaluate the effect of labeling different quantities of training samples, we set labeled training samples as 10%, 20%, 50% and 100% of the total training samples respectively. We repeat the experiments 3 times and the average classification accuracy is reported. The result is shown in Figure 4.4. We can see that our feature selection method is effective in all cases compared to the baseline of no feature selection strategy. This gives the intuition for the necessity of feature selection.

Moreover, to better understand the influence of the unlabeled data in our proposed problem, we conduct the experiments with 10% labeled (L) and 10% labeled (L) + 90% Unlabeled (U) as shown in Table 4.6. From Table 4.6, we observe that considering both labeled and unlabeled data as in our model outperform the model without unlabeled data since our model could explore the useful information from unlabeled data.

Table 4.6: Influence of the Unlabeled Data

	10% L	10% L + 90% U	20% L	20% L + 80% U	50% L	50% L + 50% U
Youtube Liu *et al.* [2009]	34.3%	37.5%	36.9%	39.7%	37.6%	41.1%
Kodak Yanagawa *et al.* [2008]	44.9%	49.3%	49.1%	53.7%	51.8%	55.6%
HumanEva Sigal *et al.* [2010]	90.1%	94.7%	94.1%	97.2%	95.3%	98.7%
MIR FLICKR Huiskes & Lew [2008]	48.8%	53.0%	51.5%	56.7%	54.3%	58.2%
COIL-20 Nene *et al.* [1996]	79.9%	82.2%	84.1%	86.9%	86.8%	89.5%
COREL-50 Hoi *et al.* [2008]	40.9%	44.3%	43.7%	46.8%	44.3%	48.7%

4.3.9 Local Sets Analysis

Figure 4.5 shows the different classification results on different datasets when considering different numbers of neighbors in our model. We find that our model achieves the best performance most of time when we consider a local structure with N = 7 neighbors. Considering too many neighbors will introduce irrelevant information in the model and considering not enough neighbors will lose the important local information for the model.

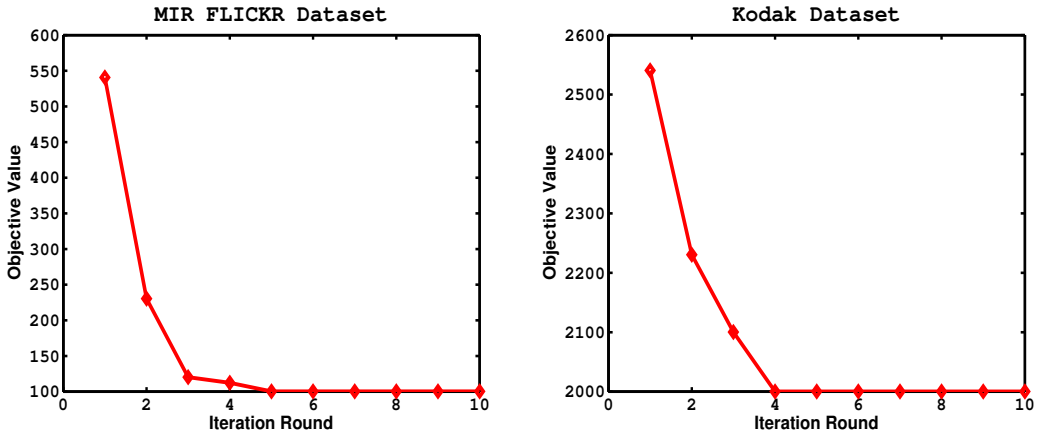
4.3.10 Convergence Analysis and Computational Cost

The proposed iterative approach monotonically decreases the objective function value in Eqn.(6) until convergence. Figure 4.6 shows the convergence curves of our algorithm on MIR FLICKR dataset and Kodak datasets. It can be observed that the objective function value converges quickly and our approach usually converges after 5 iterations at most (precision = 0.001).

Regarding the computational cost of our proposed algorithm, we train our model for MIR FLICKR dataset with 25000 samples in 10 minutes without cross-validataion on a desktop computer with Intel(R) Xeon(R) CPU E5-2620 0 @ 2.00GHz processor. This means that our algorithm would be scalable for large-scale problems.

4.4 Conclusions

In this paper, we have proposed a novel feature selection method for different multimedia applications, *i.e.*, image and video annotation and 3D motion data analysis. Our method proposes two advances over the state of the art: 1) the $\ell_{2,p}$-norm based



(a) (b)

Figure 4.6: Convergence of our algorithm on (a) MIR FLICKR dataset and (b) Kodak dataset.

sparse model; 2) the exploitation of both the global and local structures of data distribution. By using the $\ell_{2,p}$-norm based sparse model, our method is able to jointly select features across all data points and has more flexibility in choosing the discriminating subset from the original features. Meanwhile, by considering both the global and local structures of data distribution, the feature selection can be boosted as the two structures are both critical in classification tasks. Experimental results performed on real-world image and video datasets show the efficacy of our feature selection method compared to several state-of-the-art feature selection methods.

Chapter 5

Semantic Dictionary Learning for Complex Event Detection

Complex event detection is a retrieval task with the goal of finding videos of a particular event in a large scale internet video archive, given example videos and text descriptions. Nowadays, different multimodal fusion schemes of low-level and high-level features are extensively investigated and evaluated for the event detection task. Dictionary learning is a data-driven approach which learns the atom representation for a specific data source and has recently achieved great sucess in different computer vision tasks. In this paper, we firstly investigate the possibility of automatically selecting semantic meaningful concepts for the event detection task based on both the events-kit text descriptions and the concepts' high-level feature descriptions. Then we learn a semantic-oriented dictionary representation for each event based on the selected semantic concepts. To do this, we leverage training samples of selected concepts from the Semantic Indexing (SIN) dataset into a novel jointly supervised multi-task dictionary learning framework. Extensive experimental results on TRECVID Multimedia Event Detection (MED) dataset show that our proposed method outperforms the state-of-the-art methods by up to 8%.

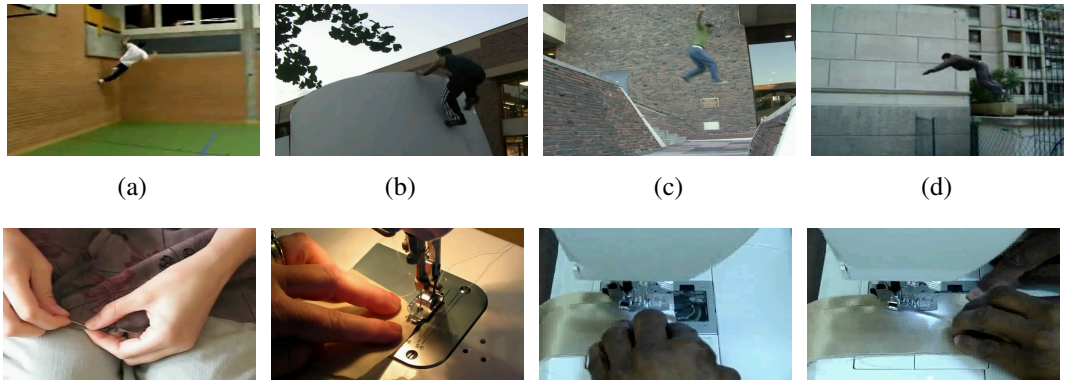
5.1 Introduction

During the past decade, due to the exponential growth of the user generated videos and the prevailing videos sharing communities such as YouTube, Hulu, *etc.*, automatic detection and retrieval of complex events in unconstrained videos has received much attention in the research community. Human action recognition from videos has been well studied Weinland *et al.* [2011] in computer vision area during the past few years. However, atomic actions such as 'running', 'jumping', *etc.* are too primitive to be used for the internet searching problem due to the complexity of internet unconstrained videos. If we consider the event detection from unconstrained recording conditions of web videos, some basic questions that need to be answered are *i.e.* 'what is an event?' or 'what defines an event in video?'. We looked up into the dictionary and saw that an 'event' refers to an observable occurrence that interests users and is found in specific scenes and is characterized by the subjects and objects, *i.e.* 'Changing vehicle unstuck', 'Making sandwich', 'Flash mob gathering'.

Complex event detection is a retrieval task NIST [2013] with the goal of detecting videos of a particular event in a large scale internet video archive, given an event-kit. An event-kit consists of example videos and text descriptions of the event. The ultimate goal for event detection is that the event engine is capable of retrieving relevant videos addressing the event of interest when a user describes a completely new event in a few sentences.

Compared with traditional concept analysis Luo *et al.* [2008]; Snoek *et al.* [2006], complex event detection is a more challenging task due to its dynamic attributes and semantic richness. For example, the event of 'Working on a sewing project' consists of a combination of several concepts such as 'sewing machine', 'people' and 'hand' together with the action 'sewing' within a longer video sequence. Figure 5.1 shows a couple of example snapshots of the videos from the event 'Parkour' and the event 'Working on a sewing project' which are defined by TRECVID Multimedia Event Detection 2011 task.

Traditional approaches for event detection rely on fusing multiple low-level features classification outputs, *i.e.* SIFT Lowe [2004], STIP Laptev [2005], MOSIFT Chen & Hauptmann [2009]. Recently, representing videos using high-level features, such as concept detectors Snoek & Smeulders [2010], appears promising for the event



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 5.1: Example snapshots of the videos from (a-d) the event 'Parkour' and (e-h) the event 'Working on a sewing project' defined by TRECVID Multimedia Event Detection 2011 task.

detection task. However, the state-of-the-art concept detector based approaches are not considering which concepts should be included in the training concept list Snoek & Smeulders [2010]. There are lots of redundant concepts in the concept list for the vocabulary construction. For example, it is highly impossible for concepts, *e.g.* 'cows', 'football' which are in the Sematic Indexing (SIN) concept list SIN [2013] to help detect certain event such as 'Landing fish' and 'Working on a sewing project'. Therefore, removing the uncorrelated concepts from the vocabulary construction might boost the event detection performance.

In this chapter, we investigate how to learn a semantic-driven representation for event detection. There are several important issues to be considered to accomplish this goal. Firstly, which concepts should be included in the learning framework. Since we want to learn a semantic-oriented dictionary representation for each event, useful concepts to be selected for each event in the learning framework are the key issue. This raises the problem of how to select necessary and meaningful concepts from a large pool of concepts for each event. Secondly, how we design a dictionary learning framework which can seamlessly learn from both the low-level features and the high-level concepts.

To facilitate reading, we first describe some abbreviations used in the paper. SIN stands for Sematic Indexing which is a dataset SIN [2013] containing 346 different categories (concepts) of images, such as car, adult, *etc*. SIN-MED stands for the high-

level concept features using the SIN concept list representing each MED video by a 346-dimensional feature (each dimension represents a concept).

Intuitively, it is highly expected to hit the right answer when we decrease the numbers of concepts in a dictionary. The overview of our framework is shown in Fig.5.2. Firstly, we automatically select semantic meaningful concepts for each MED event based on both the MED events-kit text descriptions and SIN-MED concept high-level feature representations. Then we leverage training samples of selected concepts from the SIN dataset into a jointly supervised multi-task dictionary learning framework. A semantic meaningful dictionary is learned through embedding the feature representation of original datasets (both MED dataset and SIN dataset) into a hidden shared subspace. To facilitate the detection tasks, we add label information in the learning framework to facilitate the semantic dictionary learning process. Therefore, the learned sparse codes have discriminative information and could be directly used for classification. Moreover, a novel ℓ_p-norm multi-task dictionary learning is proposed to strengthen the flexibility of the traditional ℓ_1-norm dictionary learning problem.

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

- We present one of the first works to make a comprehensive evaluation for automatic concept selection for event detection;
- We are the first to propose the semantic-oriented dictionary learning for event detection;
- We firstly construct a supervised multi-task dictionary learning framework which is capable of learning a semantic-oriented dictionary via leveraging information from selected semantic concepts;
- We propose a novel ℓ_p-norm multi-task dictionary learning framework which is more flexible than the traditional ℓ_1-norm dictionary learning problem;
- The proposed learning framework is a generic one which can be generalized into many computer vision and pattern recognition problems.

The chapter is organized as follows. Section 2 reviews related work from perspectives of event detection, dictionary learning and multi-task learning. Section 3

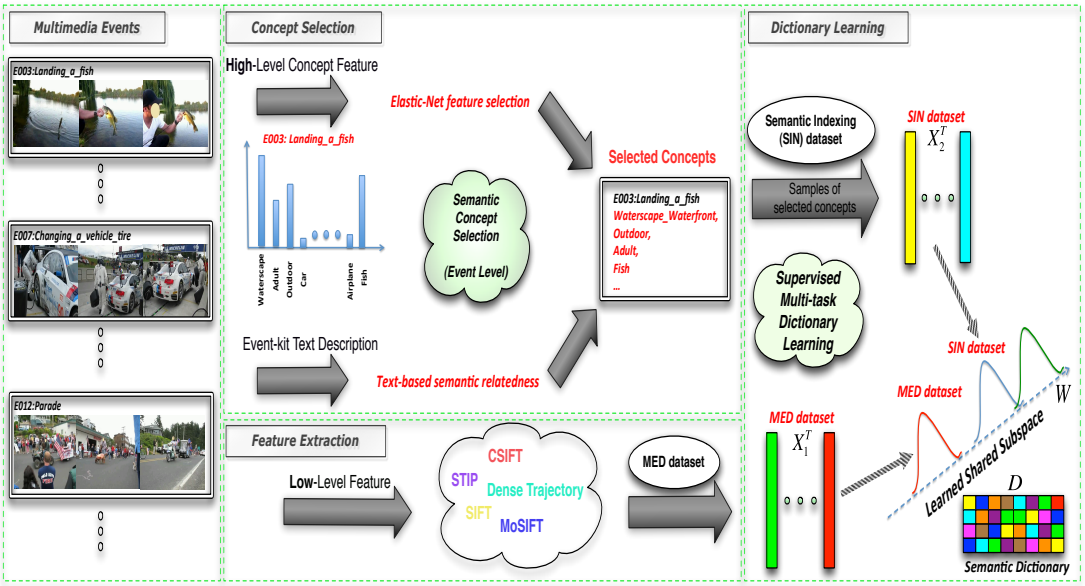


Figure 5.2: Illustration of our semantic dictionary learning framework for event detection. *Left*: Different video events. *Middle Top*: Semantic concept selection for each event based on both the MED events-kit text descriptions and SIN-MED concept high-level feature descriptions. *Middle Bottom*: Different types of low-level features extrection for events. *Right*: Supervised multi-task dictionary learning.

describes the details of our semantic concept selection strategy. Multi-task dictionary learning are described in details in Section 4. Discussion and evaluation of the proposed semantic dictionary learning for complex event detection are presented in Sections 5. We then conclude in Section 6.

5.2 Related Work

To highlight our research contributions, we now review related work on (a) Event Detection, (b) Dictionary Learning and (c) Multi-task Learning.

5.2.1 Event Detection

During the past decade, the Internet has witnessed an explosion of multimedia content. Understanding the inherent meaning of a piece of video is a difficult and useful task

for artificial intelligence.

The study of event detection first emerged in *structured* scenarios, *e.g.*, surveillance videos, sports and news videos. For example, in Adam *et al.* [2008], a robust real-time detection method using multiple fixed-location monitors was introduced to detect unusual events in surveillance videos. Sadlier & O′Connor [2005] proposed to use audio-visual features and support vector machine to detect events in field sports videos. Xu *et al.* [2006] presented a novel approach for event detection from the live sports game using webcasting text and broadcast videos. Wang *et al.* [2008] developed a multi-resolution bootstrapping framework for concept detection in news videos by exploring knowledge of sub-domain.

With the success of event detection in those structured videos, complex event detection from general *unconstrained* videos, such as those obtained from internet video sharing web sites like YouTube, has been receiving increasing attention in recent years. Unlike traditional action recognition of atomic actions, such as 'walking' or 'jumping' from videos, complex event detection aims to detect more complex events such as 'Birthday party', 'Attempting board trick', 'Changing a vehicle tire', *etc*. Tamrakar *et al.* [2012]; Yu *et al.* [2012] evaluated different low-level appearance as well as spatio-temporal features, appropriately quantized and aggregated into Bag-of-Words (BoW) descriptors for NIST TRECVID Multimedia Event Detection. Jiang *et al.* [2012] proposed a method for high-level and low-level features fusion based on collective classification from three steps which are training a classifier from low-level features, encoding high-level features into graphs, and diffusing the scores on the established graph to obtain the final prediction. Natarajan *et al.* [2012] evaluated a large set of low-level audio and visual features as well as high-level information from object detection, speech and video text OCR for event detection. They combined multiple features using a multi-stage feature fusion strategy with feature level early fusion using multiple kernel learning (MKL) and score level fusion using Bayesian model combination (BayCom) and weighted average fusion using video specific weights. Tang *et al.* [2012] tackled the problem of understanding the temporal structure of complex events in highly varying videos obtained from the Internet. A conditional model was trained in a max-margin framework that was able to automatically discover discriminative and interesting segments of video, while simultaneously achieving competitive accuracies on difficult detection and recognition tasks. Vahdat *et al.* [2013] presented a com-

positional model for event detection that leveraged a novel multiple kernel learning algorithm that incorporated structured latent variables.

Recently, representing video in terms of multi-model low-level features, *e.g.* SIFT Lowe [2004], STIP Laptev [2005], Dense Trajectory Wang *et al.* [2011], Mel-Frequency Cepstral Coefficients (MFCC) Davis & Mermelstein [1980], Automatic Speech Recognition (ASR), Optical Character Recognition (OCR), combined with early or late fusion schemes is the state-of-the-art Lan *et al.* [2012] for event detection. Despite of their good performance, low-level features are incapable of capturing the inherent semantic information in an event. Comparatively, high-level concept features were shown to be promising for event detection Snoek & Smeulders [2010]. High-level concept representation approaches become available nowadays due to the availability of large labeled training collections such as ImageNet Berg *et al.* [2011] and TRECVID Smeaton *et al.* [2006]. However, currently there are still few research works on how to automatically select useful concepts for the event detection.

5.2.2 Dictionary Learning

Dictionary Learning (also called Sparse Coding) has been 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. Dictionary learning has been successfully applied to a variety of problems in computer vision analysis recently. Yang *et al.* [2009] proposed a spatial pyramid matching approach based on SIFT sparse codes for *image classification*. The method used selective sparse coding instead of traditional vector quantization to extract salient properties of appearance descriptors of local image patches. Elad & Aharon [2006] addressed the *image denoising* problem, where zero-mean white and homogeneous Gaussian additive noise was to be removed from a given image. The approach taken was based on sparse and redundant representations over trained dictionaries. Using the K-SVD algorithm, the authors obtained a dictionary that described the image content effectively. For *image segmentation* problem, Mairal *et al.* [2008] proposed an energy formulation with both sparse reconstruction and class discrimination components, jointly optimized during dictionary learning. The approach improved over the state of the art in image segmentation experiments. Mairal *et al.* [2009b] proposed a new image model that combined the non-local means and sparse coding

approaches to *image restoration* into a unified framework where similar patches were decomposed using similar sparsity patterns.

Different optimization algorithms have also been proposed to solve dictionary learning problems. Aharon *et al.* [2006] proposed a novel K-SVD algorithm for adapting dictionaries in order to achieve sparse signal representations. K-SVD is an iterative method that alternates between sparse coding of the examples based on the current dictionary and a process of updating the dictionary atoms to better fit the data. The update of the dictionary columns was combined with an update of the sparse representations, thereby accelerating the convergence. Lee *et al.* [2006] presented efficient sparse coding algorithms that were based on iteratively solving two convex optimization problems: an ℓ_1-regularized least squares problem and an ℓ_2-constrained least squares problem. To learn a discriminative dictionary for sparse coding, a label consistent K-SVD (LC-KSVD) algorithm was proposed in Jiang *et al.* [2011]. In addition to using class labels of training data, the authors also associated label information with each dictionary item (columns of the dictionary matrix) to enforce discriminability in sparse codes during the dictionary learning process. More specifically, a new label consistent constraint was introduced and combined with the reconstruction error and the classification error to form a unified objective function. To effectively handle very large training sets and dynamic training data changing over time, Mairal *et al.* [2009a] proposed an online optimization algorithm for dictionary learning, based on stochastic approximations, which scaled up gracefully to large datasets with millions of training samples.

However, so far as we know, there is no research work on how to learn the dictionary representation at the event level for event detection and there is no research work on how to simultaneously leverage the semantic information to learn a semantic-oriented dictionary.

5.2.3 Multi-task Learning

Multi-task learning Evgeniou & Pontil [2004] methods aim 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 account for task relationships. The goal of multi-task learning is to improve the performance of learning algorithms by learning

classifiers for multiple tasks jointly. This works particularly well if these tasks have some commonality and are generally slightly under-sampled.

To capture the task relatedness from multiple related tasks is to constrain all models to share a common set of features. This motivates the group sparsity, *i.e.* the ℓ_1/ℓ_2-norm regularized learning Argyriou *et al.* [2007, 2008]. The joint feature learning using ℓ_1/ℓ_q-norm regularization performs well in ideal cases. In practical applications, however, simply using the ℓ_1/ℓ_q-norm regularization may not be effective for dealing with dirty data which may not fall into a single structure. To this end, the dirty model for multi-task learning was proposed in Jalali *et al.* [2010]. Another way to capture the task relationship is to constrain the models from different tasks to share a low-dimensional subspace by the trace norm Ji & Ye [2009]. The assumption that all models share a common low-dimensional subspace is restrictive in some applications. To this end, an extension that learns incoherent sparse and low-rank patterns simultaneously was proposed in Chen *et al.* [2010].

Many multi-task learning algorithms assume that all learning tasks are related. In practical applications, the tasks may exhibit a more sophisticated group structure where the models of tasks from the same group are closer to each other than those from a different group. There have been many works along this line of research Jacob *et al.* [2008]; Zhang & Yeung [2010]; Zhou *et al.* [2011a], known as clustered multi-task learning (CMTL). Moreover, most multi-task learning formulations assume that all tasks are relevant, which is however not the case in many real-world applications. Robust multi-task learning (RMTL) is aimed at identifying irrelevant (outlier) tasks when learning from multiple tasks Chen *et al.* [2011].

However, there is little work on multi-task learning used for dictionary learning problem. The only related theoretical work is that in Maurer *et al.* [2013], where only theoretical bounds are provided on evaluating the generalization error of dictionary learning for multi-task learning and transfer learning. Multi-task learning has received considerable attention in the computer vision community and has been successfully applied to many computer vision problems, such as image classification Yuan & Yan [2010], image annotation Quattoni *et al.* [2009] and visual tracking Zhang *et al.* [2012]. However, to our knowledge, no previous works have considered the problem of complex event detection.

5.3 Semantic Concept Selection

The concepts, which are related to objects, actions, scences, attributes, *etc.* are usually the basic elements for the description of an event. For example, as shown in Fig.5.3, concepts such as 'person', 'waterscape', 'hand' and 'fish' are the most important elements for the event 'Landing_a_fish'. However, concepts such as 'cake', 'firework', 'dog' and 'car' are not relevant to the event. In this section, we discuss which relevant concepts should be selected for the specific event for the semantic dictionary learning procedure. Both of human *linguistic knowledge* from MED event-kit text description and *visual high-level* semantic representation of each event are considered in our semantic concept selection strategy.

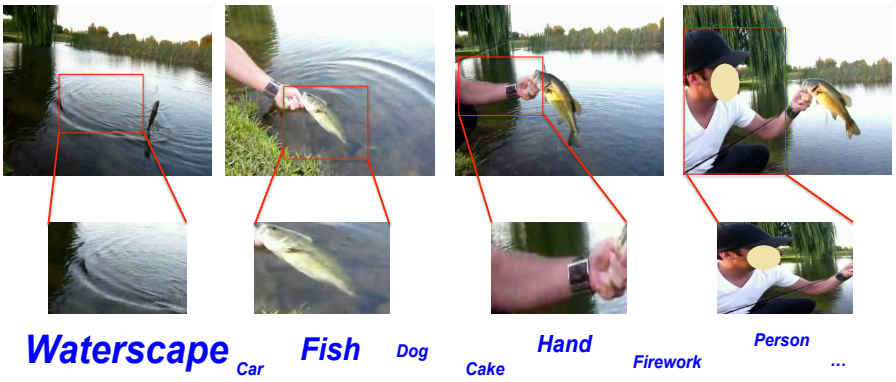


Figure 5.3: Concepts such as 'person', 'waterscape', 'hand' and 'fish' are the most important elements for the event 'Landing_a_fish'. Concepts such as 'cake', 'firework', 'dog' and 'car' are NOT relevant to the event.

5.3.1 Linguistic: Text-based Semantic Relatedness

The most widely used resources in Natural Language Processing (NLP) to calulate the semantic relateness of concepts are WordNet Fellbaum [1998], Wikipedia Strube & Ponzetto [2006] and the Word Wide Web. In this paper, we explore the semantic similarity between every term in the event-kit text description provided by NIST NIST [2013] and SIN 346 visual concept names SIN [2013] based on WordNet. There are detailed event-kit text descriptions for each MED event provided by NIST NIST [2013]. Fig.5.4 shows an example of an MED event-kit text description for 'E007:

Changing_a_vehicle_tire'.

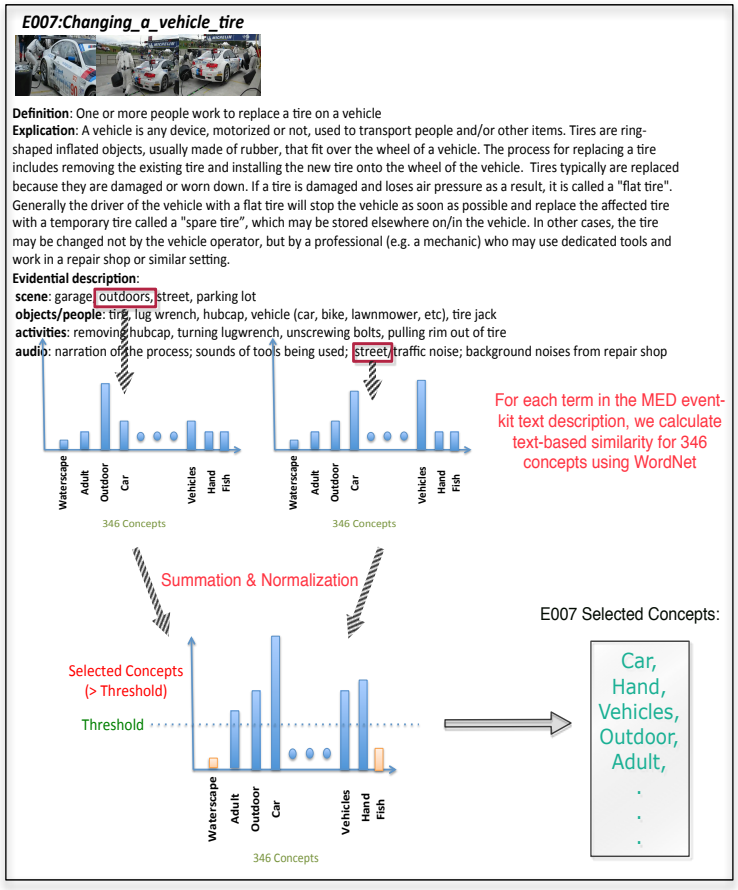


Figure 5.4: Linguistic-based concept selection strategy with an example of 'E007: Changing_a_vehicle_tire' in MED event-kit text description and a corresponding example video provided by NIST.

As illustrated in Fig.5.4, we calculate the similarity between each term in event-kit text descriptions and the SIN 346 visual concept names based on the similarity measurement proposed in Lin [1998] which defines the similarity of two words w_{1i} and w_{2j} as $sim(w_{1i}, w_{2j}) = \frac{2\ \pi(l_{cs})}{\pi(w_{1i})+\pi(w_{2j})}$, where $w_{1i} \in \{\text{event-kit text descriptions}\}$ and $i = 1, \ldots, N^{event\_kit}$, $w_{2j} \in \{\text{SIN visual concept names}\}$ and $j = 1, \ldots, 346$. l_{cs} denotes the lowest common subsumer of the two words in the WordNet hierarchy. π denotes the information content of a word and is computed as $\pi(w) = log\ p(w)$, where $p(w)$ is the probability of encountering an instance of w in a corpus. The probability $p(w)$ can be estimated from the relative corpus frequency of w and the probablities of

all words that w subsumes Resnik [1995]. In this way, we expect to capture the semantic similarity between subjects (*e.g.* human, crowd) and objects (*e.g.* animal, vehicle) effectively based on WordNet hierarchy. Finally, we construct a 346-dimensional event-level feature vector representation for each event (each dimension corresponds to an SIN visual concept name) using the MED event-kit text description *only* from linguistic knowledge. We do not use any visual information to generate the event concept representation here. A threshold is set ($thr = 0.5$ in our experiments) to select useful concepts into our final semantic concept list.

5.3.2 Visual High-level Representation: Elastic-Net Feature Selection

Low-level features are widely used for representing videos, however, they are incapable of providing insight on understanding the semantic structure underlying a video. Concept detectors provide a high-level semantic representation for videos with complicated contents, which inclines to benefit for developing powerful retrieval or filtering systems for consumer media Snoek & Smeulders [2010]. In our case, we extract semantic indexing (SIN-MED) features of a video to predict the 346 semantic concepts existing in its keyframes. SIFT is used to describe the information of images. Bag-of-words SIFT is used to train a model for each concept. Once we have the prediction score of each concept on each keyframe, the keyframe can be represented as a 346-dimensional feature indicating the determined concept probabilities. The video-level SIN-MED feature is computed as the average of keyframe-level SIN-MED feature.

To obtain concept representations for each event, we adopt the Elastic-Net Zou & Hastie [2005] feature selection as illustrated in Fig.5.5, given the intuition that the learner generally would like to choose the most representative SIN-MED feature dimensions (concepts) to differentiate events. Elastic-Net is formulated as follows:

$$\min_{\mathbf{u}} \|\mathbf{l} - \mathbf{F}\mathbf{u}\|^2 + \alpha_1 \|\mathbf{u}\|_1 + \alpha_2 \|\mathbf{u}\|^2$$

where $\mathbf{l} = \{0, 1\}^n \in I\!R^n$ indicates the event labels, $\mathbf{F} \in I\!R^{n \times b}$ is the SIN-MED feature matrix (n is the number of samples and b is the SIN-MED feature dimension) and $\mathbf{u} \in I\!R^b$ is the parameter to be optimized. Each dimension of $\mathbf{u}$ corresponds to one semantic

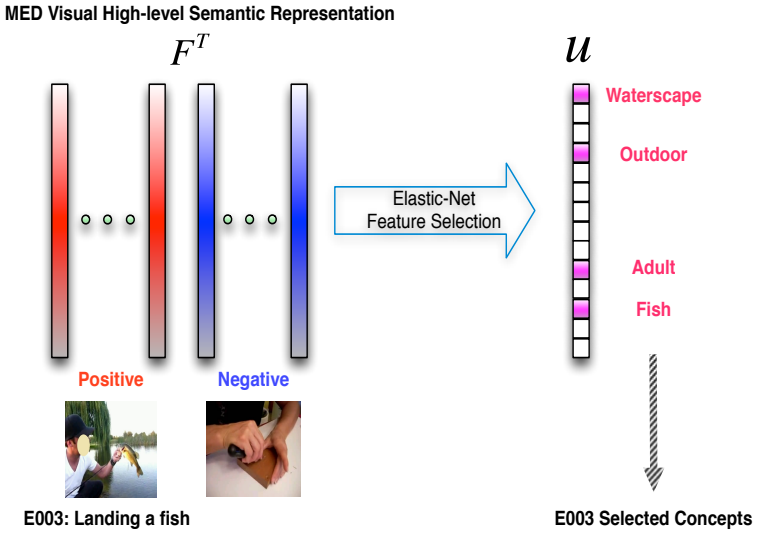


Figure 5.5: Visual high-level semantic representation with elastic-net feature selection.

concept if $\mathbf{F}$ is the high-level SIN-MED feature. α_1 and α_2 are the regularization parameters. We use Elastic-Net instead of LASSO because the concepts in the SIN concept lists SIN [2013] are highly correlated as shown in Fig.5.6. While LASSO (when $\alpha_2 = 0$) tends to select only a small number of variables from a group and ignore the others, elastic-net overcomes the limitation of LASSO and adds a quadratic term $\|\mathbf{u}\|^2$ to the penalty. We can adjust the value of α_1 value to control the sparsity degree, *i.e.*, how many semantic concepts are selected in our problem. The concepts to be selected are the corresponding dimensions with non-zero vaules of $\mathbf{u}$.

To sum up our semantic concept selection strategy, we combine the semantic concepts selected from both human linguistic as described in section 5.3.1 and visual high-level semantic representation as described in section 5.3.2 to form the final list of selected concepts for each MED event.

5.4 Semantic Dictionary Learning

After we select semantic meaningful concepts, we can then leverage training samples of selected concepts from the SIN dataset into a jointly supervised multi-task dictionary learning framework. In this section, we investigate how to learn a semantic-oriented dictionary for each event.

5.4.1 Multi-task Dictionary Learning

Given K tasks (*e.g.* $K = 2$ in our case, one task is the MED dataset and the other task is the subset of SIN dataset where samples are from specified selected concepts for each event), each task consists of data samples denoted by $\mathbf{X_k} = \{\mathbf{x_k^1}, \mathbf{x_k^2}, ..., \mathbf{x_k^{n_k}}\} \in I\!R^{n_k \times d}, (k = 1, ..., K)$, where $\mathbf{x_k^i} \in I\!R^d$ 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 across all tasks, obtained by an orthonormal projection $\mathbf{W} \in I\!R^{d \times s}$, 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 datasets via coding analysis. We consider the following optimization problem:

$$\begin{aligned} \min_{\mathbf{D_k}, \mathbf{C_k}, \mathbf{W}, \mathbf{D}} & \sum_{k=1}^{K} \|\mathbf{X_k} - \mathbf{C_k}\mathbf{D_k}\|_F^2 + \lambda_1 \sum_{k=1}^{K} \|\mathbf{C_k}\|_1 \\ & + \lambda_2 \sum_{k=1}^{K} \|\mathbf{X_k}\mathbf{W} - \mathbf{C_k}\mathbf{D}\|_F^2 \end{aligned} \tag{5.1}$$

$$s.t. \quad \begin{cases} \mathbf{W^T W} = \mathbf{I} \\ (\mathbf{D_k})_{\mathbf{j}\cdot} (\mathbf{D_k})_{\mathbf{j}\cdot}^{\mathbf{T}} \leq 1, \quad \forall j = 1, ..., l \\ \mathbf{D_{j\cdot}} \mathbf{D_{j\cdot}^T} \leq 1, \quad \forall j = 1, ..., l \end{cases}$$

where $\mathbf{D_k} \in I\!R^{l \times d}$ is an overcomplete dictionary ($l > d$) with l prototypes of the k-th task, $(\mathbf{D_k})_{\mathbf{j}\cdot}$ in the constraints denotes the j-th row of $\mathbf{D_k}$, $\mathbf{C_k} \in I\!R^{n_k \times l}$ are the sparse representation coefficients of $\mathbf{X_k}$. In the third term of Eqn.5.1, $\mathbf{X_k}$ is projected by $\mathbf{W}$ to the subspace to explore the relationship among different tasks. $\mathbf{D} \in I\!R^{l \times s}$ is the dictionary (semantic dictionary in our case) learned in the datasets shared subspace. $\mathbf{D_{j\cdot}}$ in the constraints denotes the j-th row of $\mathbf{D}$. $\mathbf{I}$ is the identity matrix. $(\cdot)^T$ denotes the transpose operator. λ_1 and λ_2 are the regularization parameters. The first constraint guarantees the learned $\mathbf{W}$ 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. Since one task in our model uses samples from the SIN dataset of selected semantic meaningful

concepts, the shared learned dictionary $\mathbf{D}$ is the semantic-oriented dictionary. When $\lambda_2 = 0$, Eqn.5.1 reduces to the traditional dictionary learning on separated tasks.

5.4.2 Supervised Multi-task Dictionary Learning

It is well-known that the traditional dictionary learning framework is not directly available for classification and the learned dictionary has merely been used for signal reconstruction Mairal *et al.* [2008]. 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 dictionary learning of Eqn.5.1 to be suitable for event detection.

Assuming that the k-th task has m_k classes, the label information of the k-th task is $\mathbf{Y_k} = \{\mathbf{y_k^1}, \mathbf{y_k^2}, ..., \mathbf{y_k^{n_k}}\} \in I\!R^{n_k \times m_k}, (k = 1, ..., K)$, $\mathbf{y_k^i} = [0, ..., 0, 1, 0, ..., 0]$ (the position of non-zero element indicates the class). $\mathbf{\Theta_k} \in I\!R^{l \times m_k}$ is the parameter of the k-th task classifier. Inspired by Zhang & Li [2010], we consider the following optimization problem:

$$\begin{aligned} \min_{\mathbf{D_k}, \mathbf{C_k}, \mathbf{\Theta_k}, \mathbf{W}, \mathbf{D}} & \sum_{k=1}^{K} \|\mathbf{X_k} - \mathbf{C_k}\mathbf{D_k}\|_F^2 + \lambda_1 \sum_{k=1}^{K} \|\mathbf{C_k}\|_1 \\ & + \lambda_2 \sum_{k=1}^{K} \|\mathbf{X_k}\mathbf{W} - \mathbf{C_k}\mathbf{D}\|_F^2 + \lambda_3 \sum_{k=1}^{K} \|\mathbf{Y_k} - \mathbf{C_k}\mathbf{\Theta_k}\|_F^2 \\ s.t. \quad & \begin{cases} \mathbf{W^T W} = \mathbf{I} \\ (\mathbf{D_k})_{\mathbf{j}\cdot} (\mathbf{D_k})_{\mathbf{j}\cdot}^{\mathbf{T}} \leq 1, \quad \forall j = 1, ..., l \\ \mathbf{D_{j\cdot}} \mathbf{D_{j\cdot}^T} \leq 1, \quad \forall j = 1, ..., l \end{cases} \end{aligned} \tag{5.2}$$

Compared with Eqn.5.1, we add the last term into Eqn.5.2 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.

5.4.3 Optimization for Eqn.5.2

To solve the proposed objective problem of Eqn.5.2, we adopt the alternating minimization algorithm to optimize it with respect to $\mathbf{D}$, $\mathbf{D_k}$, $\mathbf{C_k}$, $\mathbf{\Theta_k}$ and $\mathbf{W}$ respectively

in five steps as follows:

Step1: Fixing $\mathbf{D_k}$, $\mathbf{C_k}$, $\mathbf{W}$, $\mathbf{\Theta_k}$, Optimize $\mathbf{D}$. If we stack $\mathbf{X} = [\mathbf{X_1^T}, ..., \mathbf{X_k^T}]^\mathbf{T}$, $\mathbf{C} = [\mathbf{C_1^T}, ..., \mathbf{C_k^T}]^\mathbf{T}$, Eqn.5.2 is equivalent to:

$$\begin{aligned} &\min_{\mathbf{D}} \sum_{k=1}^{K} \|\mathbf{X_k W} - \mathbf{C_k D}\|_F^2 = \min_{\mathbf{D}} \|\mathbf{XW} - \mathbf{CD}\|_F^2 \\ &s.t. \qquad \mathbf{D_{j\cdot}} \mathbf{D_{j\cdot}^T} \leq 1, \quad \forall j = 1, ..., l \end{aligned}$$

This is equivalent to the dictionary update stage in traditional dictionary learning algorithm. We adopt the dictionary update strategy of the Algorithm 2 in Mairal *et al.* [2009a] to efficiently solve it.

Step2: Fixing $\mathbf{D}$, $\mathbf{C_k}$, $\mathbf{W}$, $\mathbf{\Theta_k}$, Optimize $\mathbf{D_k}$. Eqn.5.2 is equivalent to:

$$\begin{aligned} &\min_{\mathbf{D_k}} \|\mathbf{X_k} - \mathbf{C_k D_k}\|_F^2 \\ &s.t. \quad (\mathbf{D_k})_{\mathbf{j}\cdot} (\mathbf{D_k})_{\mathbf{j}\cdot}^\mathbf{T} \leq 1, \quad \forall j = 1, ..., l \end{aligned}$$

This is also equivalent to the dictionary update stage in traditional dictionary learning for k tasks. We adopt the dictionary update strategy of the Algorithm 2 in Mairal *et al.* [2009a] to efficiently solve it.

Step3: Fixing $\mathbf{D_k}$, $\mathbf{W}$, $\mathbf{D}$, $\mathbf{\Theta_k}$, Optimize $\mathbf{C_k}$. Eqn.5.2 is equivalent to:

$$\begin{aligned} \min_{\mathbf{C_k}} \sum_{k=1}^{K} \|\mathbf{X_k} - \mathbf{C_k D_k}\|_F^2 + \lambda_1 \sum_{k=1}^{K} \|\mathbf{C_k}\|_1 \\ + \lambda_2 \sum_{k=1}^{K} \|\mathbf{X_k W} - \mathbf{C_k D}\|_F^2 + \lambda_3 \sum_{k=1}^{K} \|\mathbf{Y_k} - \mathbf{C_k \Theta_k}\|_F^2 \end{aligned}$$

This formulation can be decoupled into $(n_1 + n_2 + ... + n_k)$ distinct problems:

$$\begin{aligned} \min_{\mathbf{c_k^i}} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\left\|\mathbf{x_k^i} - \mathbf{c_k^i D_k}\right\|_2^2 + \lambda_1 \left\|\mathbf{c_k^i}\right\|_1 \\ + \lambda_2 \left\|\mathbf{x_k^i W} - \mathbf{c_k^i D}\right\|_2^2 + \lambda_3 \left\|\mathbf{y_k^i} - \mathbf{c_k^i \Theta_k}\right\|_2^2) \end{aligned}$$

We adopt the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) Beck & Teboulle [2009] to solve the problem. 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, $r(\boldsymbol{\mu})$ is convex but non-smooth. FISTA becomes a popular tool for solving many convex smooth/non-smooth problems. In our setting, we denote the smooth term part as $f(\boldsymbol{c_k^i}) = \left\|\mathbf{x_k^i} - \mathbf{c_k^i D_k}\right\|_2^2 + \lambda_2\left\|\mathbf{x_k^i W} - \mathbf{c_k^i D}\right\|_2^2 + \lambda_3\left\|\mathbf{y_k^i} - \mathbf{c_k^i \Theta_k}\right\|_2^2$ and the non-smooth term part as $g(\boldsymbol{c_k^i}) = \lambda_1\left\|\mathbf{c_k^i}\right\|_1$.

Step4: Fixing $\mathbf{D}$, $\mathbf{C_k}$, $\mathbf{W}$, $\mathbf{D_k}$, Optimize $\mathbf{\Theta_k}$. Eqn.5.2 is equivalent to:

$$\min_{\mathbf{\Theta_k}} \|\mathbf{Y_k} - \mathbf{C_k\Theta_k}\|_F^2$$

Setting $\frac{\partial}{\partial \mathbf{\Theta_k}} = 0$, we obtain $\mathbf{\Theta_k} = (\mathbf{C_k^T C_k})^{-1}\mathbf{C_k^T Y_k}$.

Step5: Fixing $\mathbf{D_k}$, $\mathbf{C_k}$, $\mathbf{D}$, $\mathbf{\Theta_k}$, Optimize $\mathbf{W}$. If we stack $\mathbf{X} = [\mathbf{X_1^T}, ..., \mathbf{X_k^T}]^\mathbf{T}$, $\mathbf{C} = [\mathbf{C_1^T}, ..., \mathbf{C_k^T}]^\mathbf{T}$, Eqn.5.2 is equivalent to:

$$\begin{aligned}&\min_{\mathbf{W}} \sum_{k=1}^{K} \|\mathbf{X_k W} - \mathbf{C_k D}\|_F^2 = \min_{\mathbf{W}} \|\mathbf{XW} - \mathbf{CD}\|_F^2 \\ &s.t. \quad \mathbf{W^T W} = \mathbf{I}\end{aligned}$$

Substituting $\mathbf{D} = (\mathbf{C^T C})^{-1}\mathbf{C^T XW}$ back into the above function, we achieve

$$\begin{aligned}&\min_{\mathbf{W}} \left\|(\mathbf{I} - \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T})\mathbf{XW}\right\|_F^2 \\ &= \min_{\mathbf{W}} tr(\mathbf{W^T X^T}(\mathbf{I} - \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T})\mathbf{XW}) \\ &s.t. \quad \mathbf{W^T W} = \mathbf{I}\end{aligned}$$

The optimal $\mathbf{W}$ is composed of eigenvectors of the matrix $\mathbf{X^T}(\mathbf{I} - \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T})\mathbf{X}$ corresponding to the s smallest eigenvalues.

We summarize our algorithm for solving Eqn.5.2 as Algorithm 5.

5.4.4 Supervised Multi-task ℓ_p-norm Dictionary Learning

For some dictionary learning problems, using non-convex ℓ_p-norm minimization ($0 \leq p < 1$) can often obtain better results than the convex ℓ_1-norm minimization. Inspired by this, we extend our supervised multi-task dictionary learning model to supervised multi-task ℓ_p-norm dictionary learning model.

Algorithm 5 Supervised Multi-task Dictionary Learning.

Input: K tasks Data $(\mathbf{X_1}, ..., \mathbf{X_k})$ and Label $(\mathbf{Y_1}, ..., \mathbf{Y_k})$; Subspace dimensionality s, Dictionary size l, Regularization parameters λ_1, λ_2, λ_3.

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

Initialize $\mathbf{W}$ using any orthonormal matrix;
Initialize $\mathbf{C_k}$ with l_2 normalized columns;
Repeat
Compute $\mathbf{D}$ using Algorithm 2 in Mairal *et al.* [2009a];
for $k = 1 : K$
    Compute $\mathbf{D_k}$ using Algorithm 2 in Mairal *et al.* [2009a];
    Adopting FISTA Beck & Teboulle [2009] to solve $\mathbf{C_k}$;
    $\mathbf{\Theta_k} = (\mathbf{C_k^T C_k})^{-1}\mathbf{C_k^T Y_k}$;
end for
Compute $\mathbf{W}$ by eigen decomposition of $\mathbf{X^T}(\mathbf{I} - \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T})\mathbf{X}$;
Until Convergence

Assuming that the k-th task has m_k classes, the label information of the k-th task is $\mathbf{Y_k} = \{\mathbf{y_k^1}, \mathbf{y_k^2}, ..., \mathbf{y_k^{n_k}}\} \in I\!R^{n_k\times m_k}, (k = 1, ..., K)$, $\mathbf{y_k^i} = [0, ..., 0, 1, 0, ..., 0]$ (the position of non-zero element indicates the class). $\mathbf{\Theta_k} \in I\!R^{l\times m_k}$ is the parameter of the k-th task classifier. We formulate our supervised multi-task ℓ_p-norm dictionary learning problem as follows:

$$\begin{aligned}\min_{\mathbf{D_k},\mathbf{C_k},\mathbf{\Theta_k},\mathbf{W},\mathbf{D}} &\sum_{k=1}^{K} \|\mathbf{X_k} - \mathbf{C_k D_k}\|_F^2 + \lambda_1 \sum_{k=1}^{K} \|\mathbf{C_k}\|_p^p \\ &+\lambda_2 \sum_{k=1}^{K} \|\mathbf{X_k W} - \mathbf{C_k D}\|_F^2 + \lambda_3 \sum_{k=1}^{K} \|\mathbf{Y_k} - \mathbf{C_k \Theta_k}\|_F^2 \\ s.t. \quad &\begin{cases} \mathbf{W^T W} = \mathbf{I} \\ (\mathbf{D_k})_{\mathbf{j}\cdot}(\mathbf{D_k})_{\mathbf{j}\cdot}^{\mathbf{T}} \le 1, \quad \forall j = 1, ..., l \\ \mathbf{D_{j\cdot}} \mathbf{D_{j\cdot}^T} \le 1, \quad \forall j = 1, ..., l \end{cases}\end{aligned} \tag{5.3}$$

Compared with Eqn.5.2, we replace the traditional sparse coding ℓ_1-norm term $\|\mathbf{C_k}\|_1$ with the more flexible ℓ_p-norm term $\|\mathbf{C_k}\|_p^p$. Since we can adjust the value of p $(0 \le p < 1)$ in our framework, our algorithm is more flexible to control the sparseness of the feature representation, thus usually resulting in better performance

than the traditional ℓ_1-norm sparse coding.

To solve the proposed problem of Eqn.5.3, we adopt the alternating minimization algorithm to optimize it with respect to $\mathbf{D}, \mathbf{D_k}, \mathbf{C_k}, \mathbf{\Theta_k}, \mathbf{W}$ respectively. The updated rules for $\mathbf{D}, \mathbf{D_k}, \mathbf{\Theta_k}, \mathbf{W}$ are the same as Eqn.5.2, the only difference exists in the updated rule of $\mathbf{C_k}$. Various algorithms have been proposed for ℓ_p-norm non-convex sparse coding Gorodnitsky & Rao [1997]; Krishnan & Fergus [2009]; She [2009]. In this paper, we adopt Generalized Iterated Shrinkage Algorithm (GISA) Zuo *et al.* [2013] to solve the proposed problem. We summarize our algorithm for solving Eqn.5.3 as Algorithm 6.

Algorithm 6 Supervised Multi-task ℓ_p-norm Dictionary Learning.

Input: K tasks Data $(\mathbf{X_1}, ..., \mathbf{X_k})$ and Label $(\mathbf{Y_1}, ..., \mathbf{Y_k})$; Subspace dimensionality s, Dictionary size l, ℓ_p-norm parameter p, Regularization parameters λ_1, λ_2, λ_3.
Output: Optimized $\mathbf{W} \in \mathbb{R}^{d\times s}$, $\mathbf{C_k} \in \mathbb{R}^{n_k\times l}$, $\mathbf{D_k} \in \mathbb{R}^{l\times d}$, $\mathbf{D} \in \mathbb{R}^{l\times s}$, $\mathbf{\Theta_k} \in \mathbb{R}^{l\times m_k}$.

Initialize $\mathbf{W}$ using any orthonormal matrix;
Initialize $\mathbf{C_k}$ with l_2 normalized columns;
Repeat
Compute $\mathbf{D}$ using Algorithm 2 in Mairal *et al.* [2009a];
for $k = 1 : K$
    Compute $\mathbf{D_k}$ using Algorithm 2 in Mairal *et al.* [2009a];
    Adopting GISA Zuo *et al.* [2013] to solve $\mathbf{C_k}$;
    $\mathbf{\Theta_k} = (\mathbf{C_k^T C_k})^{-1}\mathbf{C_k^T Y_k}$;
end for
Compute $\mathbf{W}$ by eigen decomposition of $\mathbf{X^T}(\mathbf{I} - \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T})\mathbf{X}$;
Until Convergence

After the optimized $\mathbf{\Theta}$ is obtained, the final classification of a test video can be obtained based on its sparse coefficient $\mathbf{c_k^i}$, which carries the discriminative information. We can simply apply the linear classifier $\mathbf{c_k^i \Theta_k}$ to obtain the predicted score of the video.

5.5 Experiments

In this section, we conduct extensive experiments to evaluate our proposed method for TRECVID MED task.

5.5.1 Datasets

TRECVID MED10 (P001-P003) and MED11 (E001-E015) datasets are used in our experiments. The datasets consist of 9746 videos from 18 events of interest, with 100-200 examples per event, and the rest of the videos are from the background class. The details are listed in the Table 5.1.

Table 5.1: 18 Events of TRECVID MED10 and MED11

	Event Name	Train-set Positive #	Train-set Negative #	Test-set Positive #	Test-set Negative #
P001:	**Assembling shelter**	51	3053	45	6597
P002:	**Batting a run**	54	3050	52	6590
P003:	**Making a cake**	59	3045	47	6595
E001:	**Attempting board trick**	161	2943	114	6528
E002:	**Feeding animal**	162	2942	114	6528
E003:	**Landing fish**	119	1984	85	6557
E004:	**Wedding ceremony**	125	2979	87	6555
E005:	**Working wood working project**	141	2963	99	6543
E006:	**Birthday party**	87	3017	86	6556
E007:	**Changing a vehicle tire**	56	3048	55	6587
E008:	**Flash mob gathering**	87	3017	86	6556
E009:	**Getting a vehicle unstuck**	64	3040	66	6576
E010:	**Grooming an animal**	69	3035	69	6573
E011:	**Making a sandwich**	62	3042	63	6579
E012:	**Parade**	68	3036	69	6573
E013:	**Parkour**	56	3048	55	6587
E014:	**Repairing an appliance**	62	3042	61	6581
E015:	**Working on a sewing project**	60	3044	60	6582

TRECVID Semantic Indexing Task (SIN) SIN [2013] contains annotation for 346 semantic concepts on 400,000 keyframes from web videos. 346 concepts are related to objects, actions, scences, attributes and non-visual concepts which are all the basic elements for an event, *e.g.* kitchen, boy, girl, bus. For the sake of better understanding and easy concept selection, we manually divide the 346 visual concepts into 15 groups which are listed in Table 5.2.

Table 5.2: 15 groups of TRECVID SIN 346 visual concepts (the number of concepts for each group are in parenthesis)

G1:	**Body_Parts** (8)	G2:	**Person** (14)	G3:	**Military** (76)
G4:	**Car** (27)	G5:	**Boat** (7)	G6:	**Aircraft** (10)
G7:	**Nature** (27)	G8:	**Indoor** (25)	G9:	**News** (29)
G10:	**Animal** (22)	G11:	**Urban_Scenes** (50)	G12:	**Natural_Disaster** (3)
G13:	**Election** (5)	G14:	**Sport_Activity** (33)	G15:	**Moods** (10)

5.5.2 Evaluation Metrics

MED system performance is evaluated as a binary classification system by measuring the performance of two types of errors: Missed Detection (MD) errors and False Alarm (FA) errors. The primary measure for accuracy is the probability of missed detection (the number of missed detection divided by the number of clips containing an event) and false alarms (the number of false alarms divided by the number of clips not containing the event) for the event based on the detection threshold. The three evaluation metrics we used are listed as follows:

- **Average Precision (AP)**: is a measure that combines recall and precision for ranked retrieval results. For one information need, the average precision is the mean of the precision scores after each relevant sample is retrieved. The *higher* number indicates better performance.

- **PMiss@TER=12.5**: is an official evaluation metric for event detection as defined by NIST NIST [2013]. It is defined as the point at which the ratio between the probability of Missed Detection and probability of False Alarm is 12.5:1. The *lower* number indicates better performance.

- **Normalized Detection Cost (NDC)**: is an official evaluation metric for event detection as defined by NIST NIST [2013]. It is a weighted linear combination of the system's Missed Detection and False Alarm probabilities. NDC measures the performance of a detection system in the context of an application profile

using error rate estimates calculated on a test set. The *lower* number indicates better performance.

5.5.3 Experiment Setup

There are 3104 videos used for training and 6642 videos used for testing in our experiments. We use three representative features which are SIFT, Color SIFT (CSIFT) and Motion SIFT (MOSIFT) Chen & Hauptmann [2009]. SIFT and CSIFT describe the gradient and color information of images. MOSIFT describes both the optical flow and gradient information of video clips. Finally, 768-dimensional SIFT-BoW, CSIFT-BoW, MOSIFT-BoW features are extracted respectively to represent each video. 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 $\{200, 400, 600\}$. For the experiments in the paper, we try three different dictionary sizes from $\{768, 1024, 1280\}$. To evaluate the multi-task ℓ_p-norm dictionary learning algorithm, parameter p is tuned in the range of $\{0.2, 0.4, 0.6, 0.8, 1\}$.

5.5.4 Comparison Method

We compare our semantic supervised multi-task dictionary learning with the following important baselines:

- ***SVM***: SVM is an effective tool for complex event detection and has been widely used by several research groups for TRECVID MED Lan *et al.* [2012]; Natarajan *et al.* [2012]; Yu *et al.* [2012];

- *Single Task Supervised Dictionary Learning (**ST-SDL**)*: Performing supervised dictionary learning on each task separately;

- *Pooling Tasks Supervised Dictionary Learning (**PT-SDL**)*: Performing single task supervised dictionary learning by simply aggregating data from all tasks;

- *Multiple Kernel Transfer Learning (**MKTL**)* Jie *et al.* [2011]: A method which incorporates prior features into a multiple kernel learning framework;

- *Dirty Model Multi-Task Learning (**DMMTL**)* Jalali *et al.* [2010]: A state-of-the-art multi-task learning method imposing ℓ_1/ℓ_q-norm regularization;

- *Random Concept Selection Strategy (**RCSS**)*: Performing our proposed supervised multi-task dictionary learning *without* involving concept selection strategy (leveraging random samples).

5.5.5 Results

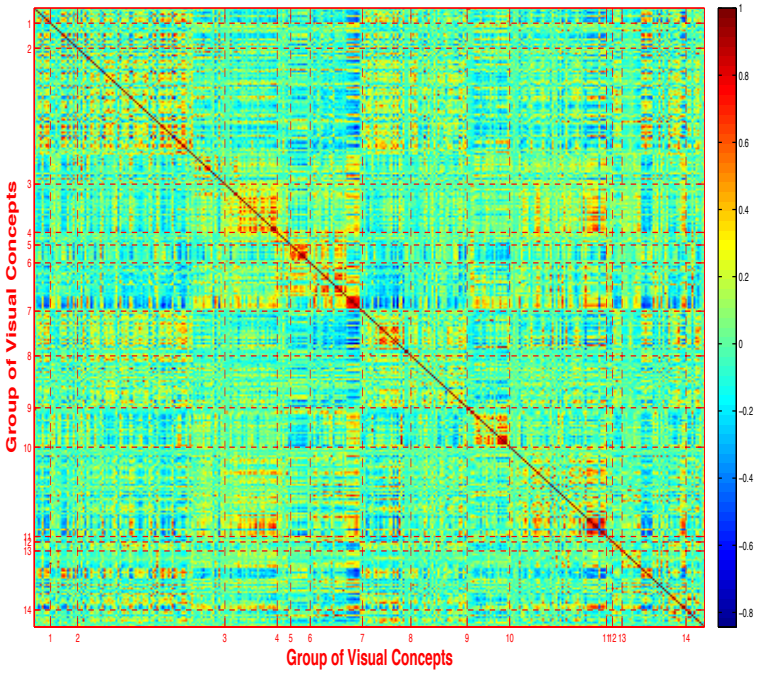


Figure 5.6: The correlation of SIN visual concepts. High correlations between contextually related clusters 'G4:car', 'G7:nature', 'G11:urban-scene' (in red) and negative correlations between contextually unrelated clusters 'G7:nature', 'G8:indoor' (in blue).

Firstly, we calculate the covariance matrix of the 346 SIN concepts, shown in

Fig.5.6, where the visual concepts are grouped and sorted the same as in Table 5.2. The convariance matrix shows the high within-cluster correlations along the diagonal direction and also relatively high correlations between contextually related clusters (in red color), such as 'G4:car', 'G7:nature' and 'G11:urban-scene'. We also observe negative correlations between contextually unrelated clusters (in blue color), such as 'G7:nature' and 'G8:indoor'. This gives us the intuition that visual concepts co-occurrence exists. Therefore, by removing redundant concepts and selecting related concepts for each event is expected to be helpful for the event detection.

Table 5.3 shows the results of top 5 ranked concepts based on our concept selection strategy for each event. Interestingly, from Table 5.3, we can observe that the concepts selected by our methods in Section 5.3 are reasonably consistent with human selections. For example, we select 'food', 'kitchen', 'hand', 'body_part', 'table' for the event 'Making a sandwich' and 'Man_made_thing', 'hand', 'table', 'furniture', 'glasses' for the event 'Repairing an appliance'.

Table 5.3: Results of top 5 ranked concepts based on the concept selection method proposed in section 5.3.

	Event Name	Rank #1	Rank #2	Rank #3	Rank #4	Rank #5
P001:	**Assembling shelter**	Outdoor	Man	Fields	Car	Chair
P002:	**Batting a run**	Fields	Athlete	Sports	Standing	Outdoor
P003:	**Making a cake**	Hand	Man	Standing	Food	Kitchen
E001:	**Attempting board trick**	Skating	Athlete	Sports	Outdoor	Standing
E002:	**Feeding animal**	Animal	Food	Cats	Hand	Dogs
E003:	**Landing fish**	Boat_Ship	Lakes	Body_Parts	Oceans	Hand
E004:	**Wedding ceremony**	Man_Wearing_A_Suit	Crowd	Dresses	Flowers	Dancing
E005:	**Working wood working project**	Construction_Worker	Adult	Man_Made_Thing	Chair	Table
E006:	**Birthday party**	3_Or_More_People	Food	Singing	Crowd	Baby
E007:	**Changing a vehicle tire**	Vehicle	Hand	Car	Truck	Motorcycle
E008:	**Flash mob gathering**	People_Marching	Crowd	Road	Walking	Outdoor
E009:	**Getting a vehicle unstuck**	Vehicle	Outdoor	Construction	Fields	Snow
E010:	**Grooming an animal**	Animal	Cats	Dogs	Birds	Mammal
E011:	**Making a sandwich**	Food	Kitchen	Hand	Body_Parts	Table
E012:	**Parade**	People_Marching	Crowd	Protest	Adult	Cheering
E013:	**Parkour**	Legs	Athlete	Sports	Urban_Park	Outdoor
E014:	**Repairing an appliance**	Man_Made_Thing	Hand	Table	Furniture	Glasses
E015:	**Working on a sewing project**	Dresses	Hand	Human_Face	Table	Indoor

Table 5.4 shows the *average* detection results of the 18 MED events for different comparison methods based on SIFT feature. We have the following observations:

- Comparing ST-SDL with SVM, we observe that performing supervised dictio-

Table 5.4: Comparision of different methods for *average* detection accuracy of SIFT feature. Better results are highlighted in bold.

Evaluation Metric	SVM	ST-SDL	PT-SDL	MKTL	DMMTL	RCSS	Proposed
AP	0.0883	0.1037	0.1336	0.1191	0.1180	0.1201	**0.1664**
PMiss@TER=12.5	0.6535	0.6447	0.6127	0.6364	0.6133	0.6221	**0.5927**
MinNDC	0.9401	0.9154	0.8644	0.8843	0.8674	0.8612	**0.8404**

nary learning is better than SVM which shows the effectiveness of dictionary learning for MED task.

- Comparing PT-SDL with ST-SDL, leveraging knowledge from the SIN dataset improves the performance for MED task.

- Our method outperforms both ST-SDL and PT-STL up by 6% and 3% in AP, which shows the benefit of the multi-task settings for our proposed problem.

- Our concept selection strategy for semantic dictionary learning performs the best for MED among all the comparison methods.

- Our proposed method outperforms 8%, 6%, 10% for AP, PMiss@TER=12.5 and MinNDC respectively compared with SVM and outperforms 4%, 3%, 2% for AP, PMiss@TER=12.5 and MinNDC compared with randomly concept selection for supervised dictionary learning. Considering the difficulty of the TRECVID MED dataset and the typically low performance of MED, the achieved results are promising for our proposed semantic dictionary learning strategy for MED.

To see the comparision for each MED event individually, we show the AP results for each MED event in Fig.5.7. Our proposed method achieves the best performance for 13 events out of a total of 18 events. Especially for event 'E004: Wedding ceremony' and 'E009: Getting a vehicle unstuck', our method outperforms 10% AP com-

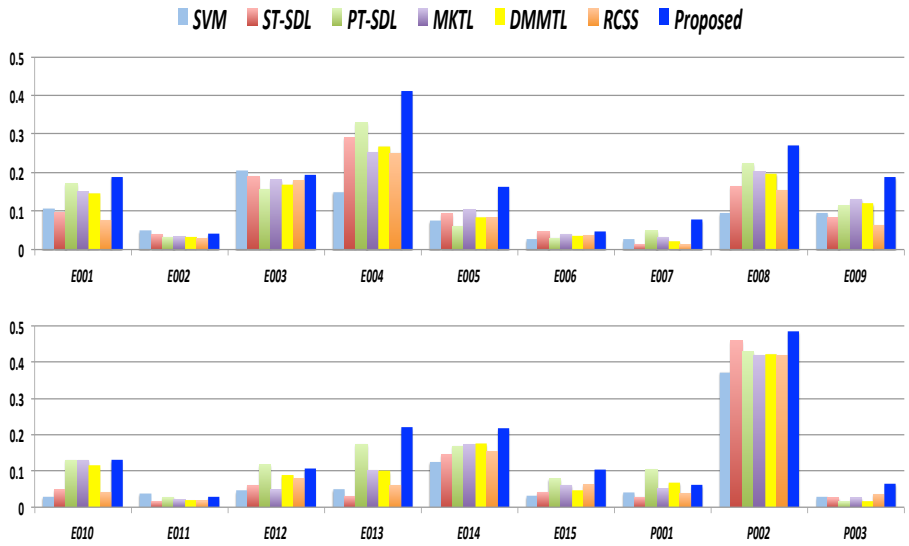


Figure 5.7: Comparision of different methods of AP performance for each MED event.

pared with the second best method. Fig.5.8 illustrates the MAP performance for different methods based on different types of SIFT, CSIFT and MOSIFT features. It can be easily observed that our proposed surpervised multi-task dictionary learning with our concept selection strategy outperforms all the other methods by at least 3% and outperforms SVM by more than 8%.

Moreover, we evaluate our proposed method with respect to different dictionary sizes and different subspace dimensionality settings based on SIFT feature. Fig.5.9(left) shows that the proposed method achieves the best MAP results when the dictionary size is 1024. Too large or too small dictionary size hurts the performance. Fig.5.9(right) shows that the best MAP result is achieved when the subspace dimensionality is 400 (dictionary size = 1024). Large or small subspace dimensionality also degrades the performance.

Finally, we also study the parameter sensitivity of the proposed method in Fig.5.10. Here, we fix $\lambda_3 = 1$ (discriminative information contribution fixed) and $p = 0.6$ and analyze the regularization parameters λ_1 and λ_2. As shown in Fig.5.10(a), 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 dictionary learning. Moverover, to understand the influence of parameter p for our proposed supervised ℓ_p-norm dictio-

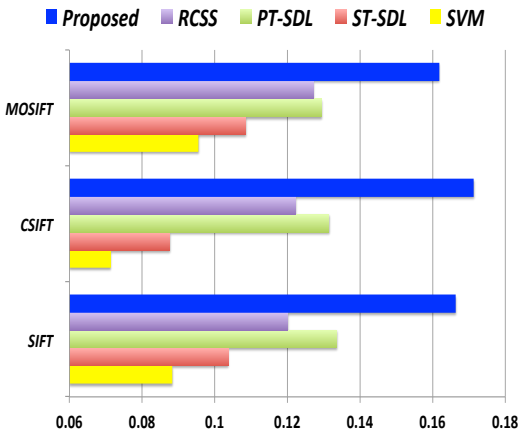


Figure 5.8: Comparision of different methods of MAP performance for different types of features.

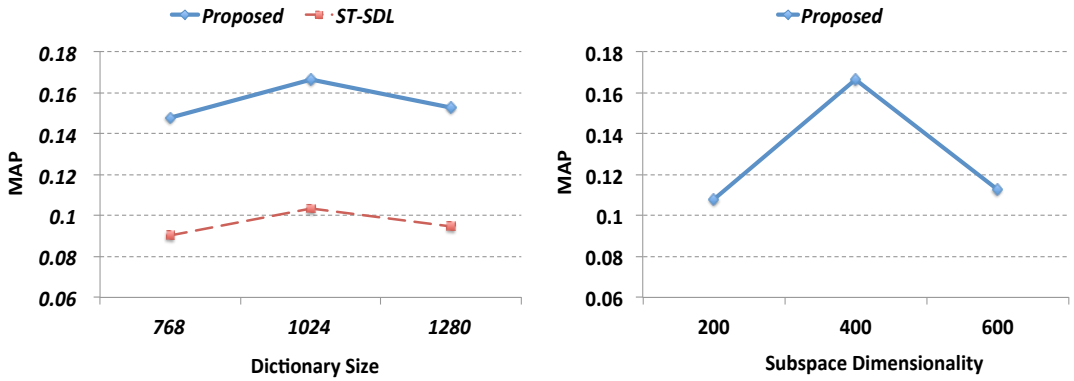
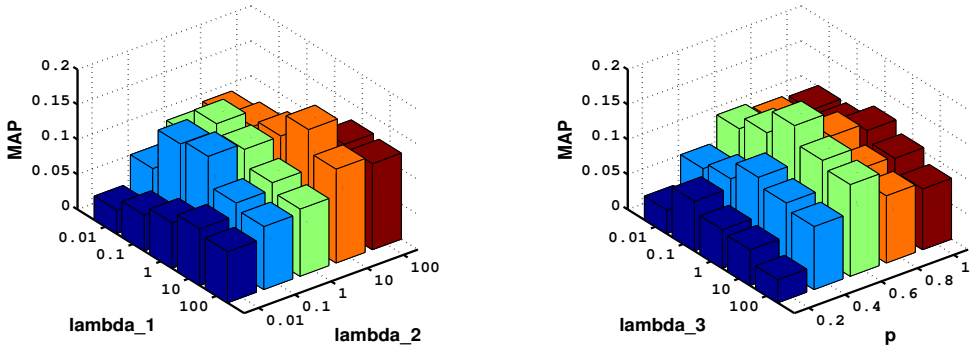


Figure 5.9: MAP performance variation w.r.t (left) different dictionary size; (right) different subspace dimensionality.

nary learning algorithm, we also perform an experiment on the parameter sensitivity. Fig.5.10(b) demonstrates that the best performance for the supervised ℓ_p-norm dictionary learning algorithm is achieved at when p = 0.6. More than 2% MAP can be achieved if we adopt a more flexible ℓ_p-norm model compared with the fixed ℓ_1-norm model. This shows the suboptimality of the traditional ℓ_1-norm sparse coding compared with the flexible ℓ_p-norm sparse coding.



(a) (b)

Figure 5.10: Sensitivity study of parameters on (a) λ_1 and λ_2 with fixed p and λ_3. (b) p and λ_3 with fixed λ_1 and λ_2.

5.6 Conclusion

This paper represents the first work to explore multi-task dictionary learning approaches for complex event detection and in particular, semantic dictionary learning, for which very few solutions have been proposed in literature. We firstly investigated the possibility to automatically select semantic meaningful concepts for a complex event detection task based on both the TRECVID MED events-kit text descriptions and the SIN-MED concept high-level feature descriptions. Then we learned a semantic-oriented dictionary representation for each event based on the selected semantic concepts. To do this, we leveraged training samples of selected concepts from the SIN dataset into a novel jointly supervised multi-task dictionary learning framework.

Extensive experimental results on the TRECVID MED dataset showed that our proposed method outperformed the state-of-the-art methods. We compared our semantic

supervised multi-task dictionary learning with several important baselines, including SVM, Single Task Supervised Dictionary Learning (ST-SDL), Pooling Tasks Supervised Dictionary Learning (PT-SDL), Random Concept Selection Strategy (RCSS). Our proposed method outperformed SVM by up to 8% MAP which showed the effectiveness of dictionary learning for TRECVID MED. More than 6% and 3% MAP was achieved respectively compared with ST-SDL and PT-SDL, which showed the advantage of multi-task setting of our proposed framework. To show the benefit of concept selection strategy, we compared RCSS to our method and showed that achieves 4% less MAP. Moreover, two state-of-the-art multi-task learning methods, Multiple Kernel Transfer Learning (MKTL) Jie *et al.* [2011] and Dirty Model Multi-Task Learning (DMMTL) Jalali *et al.* [2010], are also compared with our method.

For some sparse coding problems, non-convex ℓ_p-norm minimization ($0 \leq p < 1$) can often obtain better results than the convex ℓ_1-norm minimization. Inspired by this, we extended our supervised multi-task dictionary learning model to a supervised multi-task ℓ_p-norm dictionary learning model. We evaluated the influence of the ℓ_p-norm parameter p for our proposed problem and found that more than 2% MAP can be achieved if we adopted the more flexible ℓ_p-norm model compared with the fixed ℓ_1-norm model.

Overall, the proposed multi-task dictionary learning solutions are novel in the context of complex event detection, which is a relevant and important research problem in applications such as video understanding and surveillance. Future research involves (i) integration of knowledge from multiple sources (video, audio, text) and incorporation of kernel learning in our framework, and (ii) the use of deep structures instead of a shallow single-layer model in the proposed problem since deep learning has achieved the supreme success in many different fields of computer vision.

Chapter 6

Conclusions

In this thesis, we have addressed several computer vision problems with the focus on different applications, *i.e.*, headpose estimation, action recognition, and multimedia event detection, under multi-task learning framework.

Headpose estimation is a fundamental problem in the computer vision. Knowledge of where a person is looking as given by head pose and eye-gaze, is useful in human computer interaction as well as human behavior analysis. In Chapter 2, we proposed a novel Multi-task Learning framework (FEGA-MTL) for classifying the head pose of a person who moves freely in an environment monitored by multiple, large field-of-view surveillance cameras.

Following the progress of the headpose estimation, we have tackled the understanding of a image or a video depicting a human action. In Chapter 3, we proposed Multi-task Linear Discriminant Analysis, a novel multi-task learning framework for multi-view action recognition that allows for the sharing of discriminative Self-Similarity Matrices features among different views.

While in Chapter 3, we focus on the multi-camera setup for human action recognition, to address action recognition problem for the traditional single camera setup, in Chapter 4, we proposed a novel feature selection method using a sparse model. Different from the state of the art, our method is built upon the $l_{2,p}$-norm and simultaneously considers both the global and local (GLocal) structures of data distribution.

During the past decade, due to the exponential growth of the user generated videos and the prevailing videos sharing communities such as YouTube, Hulu, *etc.*, automatic detection and retrieval of complex events in unconstrained videos has received much

attention in the research community. In Chapter 5, we firstly investigated the possibility of automatically selecting semantic meaningful concepts for the event detection task based on both the events-kit text descriptions and the concepts high-level feature descriptions. Then we proposed a Multi-task dictionary learning framework to learn a semantic-oriented dictionary representation for each event based on the selected semantic concepts.

In summary, in this thesis we have studied different computer vision and multimedia problems under the framework of Multi-task learning. Our work suggests that the proper usage of multi-task learning and feature selection does help improve the overall understanding of computer vision and multimedia contents. Hence, in the future we will continue our research in this direction with the following possible pursuits:

- Integration of knowledge from multiple sources (video, audio, text) and incorporation of kernel learning in our multi-task learning framework;
- The use of deep structures instead of a shallow single-layer model in the proposed problem since deep learning has achieved the supreme success in many different fields of computer vision;
- With the appearing of huge amounts of data in computer vision and multimedia analysis, we will develop more scalable algorithms for Big Data analytics in the future.

References

ADAM, A., RIVLIN, E., SHIMSHONI, I. & REINITZ, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**, 555–560. 78

AHARON, M., ELAD, M. & BRUCKSTEIN, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Image Processing*, **54**, 4311–4322. 80

ARGYRIOU, A., EVGENIOU, T. & PONTIL, M. (2007). Multi-task feature learning. In *Neural Information Processing Systems*. 9, 18, 19, 25, 29, 44, 45, 46, 56, 81

ARGYRIOU, A., EVGENIOU, T. & PONTIL, M. (2008). Convex multi-task feature learning. *Machine Learning*, **73**, 243–272. 81

BECK, A. & TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, **2**, 183–202. 15, 36, 88, 90

BELHUMEOUR, P., HESPANHA, J. & KRIEGMAN, D. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**, 711–720. 29

BENFOLD, B. & REID, I. (2011). Unsupervised learning of a scene-specific coarse gaze estimator. In *IEEE International conference on Computer Vision*. 8, 12

BERG, A., DENG, J., SATHEESH, S., SU, H. & LI, F.F. (2011). Imagenet large scale visual recognition challenge. 79

BOBICK, A. & DAVIS, J. (2001). The representation and recognition of action using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 257–267. 42

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. & ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, **3**, 1–122. 16, 36, 39

CAWLEY, G.C., TALBOT, N.L.C. & GIROLAMI, M. (2006). Sparse multinomial logistic regression via bayesian L1 regularisation. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 62

CHEN, C. & ODOBEZ, J.M. (2012). We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video. In *IEEE conference on Computer Vision and Pattern Recognition*. 6, 8, 12

CHEN, J., LIU, J. & YE, J. (2010). Learning incoherent sparse and low-rank patterns from multiple tasks. In *ACM SIGKDD international conference on Knowledge discovery and data mining*. 81

CHEN, M.Y. & HAUPTMANN, A. (2009). Mosift: Recognizing human actions in surveillance videos. In *CMU Technical Report, CMU-CS-09-161*. 74, 94

CHEN, X., LIN, Q., KIM, S., CARBONELL, J. & XING, E. (2011). Smoothing proximal gradient method for general structured sparse learning. In *Uncertainity in Artifical Intelligence*. 9, 21, 29, 81

CHENG, Z., QIN, L., YE, Y., HUANG, Q. & TIAN, Q. (2012). Human daily action analysis with multi-view and color-depth data. In *European Conference on Computer Vision Workshop on Consumer Depth Cameras for Computer Vision*. 26, 39, 42, 46, 48

DALAL, N. & TRIGGS, B. (2005). Histograms of oriented gradients for human detection. In *IEEE conference on Computer Vision and Pattern Recognition*. 6

DAVIS, S. & MERMELSTEIN, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28**, 357–366. 79

DUDA, R., HART, P. & STORK, D. (2001). *Pattern Classification*. John Wiley & Sons, New York, 2nd edn. 33, 34, 51, 62, 66

EFROS, A.A., BERG, A.C., BERG, E.C., MORI, G. & MALIK, J. (2003). Recognizing action at a distance. In *IEEE International conference on Computer Vision*. 24

ELAD, M. & AHARON, M. (2006). Image denoising via sparse and redundant representation over learned dictionaries. *IEEE Transactions on Image Processing*, **15**, 3736–3745. 79

EVGENIOU, T. & PONTIL, M. (2004). Regularized multi-task learning. In *ACM SIGKDD international conference on Knowledge discovery and data mining*. 1, 9, 25, 29, 80

FARHADI, A. & TABRIZI, M.K. (2008). Learning to recognize activities from the wrong view point. In *European Conference on Computer Vision*. 25, 28, 47

FARHADI, A., TABRIZI, M.K., ENDRES, I. & FORSYTH, D.A. (2009). A latent model of discriminative aspect. In *IEEE International conference on Computer Vision*. 28

FELLBAUM, C. (1998). Wordnet: An electronical lexical database. In *The MIT Press, Cambridge, MA*. 82

FUKUNAGA, K. (1990). Introduction to statistical pattern classification. In *USA: Academic Press*. 32, 33

GAO, Y., WANG, M., ZHA, Z.J., SHEN, J., LI, X. & WU, X. (2013). Visual-textual joint relevance learning for tag-based social image search. *IEEE Transactions on Image Processing*, **22**, 363–376. 53

GONG, P., YE, J. & ZHANG, C. (2012). Robust multi-task feature learning. In *Conference on Knowledge Discovery and Data Mining (SIGKDD)*. 9, 18, 29

GORODNITSKY, I.F. & RAO, B.D. (1997). A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, **45**, 600–616. 91

GRUNDMANN, M., MEIER, F. & ESSA, I. (2008). 3d shape context and distance transform for action recognition. In *International Conference on Pattern Recognition*. 24

GUO, K., ISHWAR, P. & KONRAD, J. (2013). Action recognition from video using feature covariance matrices. *IEEE Transactions on Image Processing*, **22**, 2479–2494. 24

HAN, Y., WU, F., JIA, J., ZHUANG, Y. & YU, B. (2010). Multi-task sparse discriminant analysis (mtsda) with overlapping categories. In *Conference on Artificial Intelligence (AAAI)*. 29

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2001). The elements of statistical learning: Data mining, inference, and prediction. In *Springer*. 34

HE, X. & NIYOGI, P. (2003). Locality preserving projections. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 53

HOI, S.C., JIN, R., ZHU, J. & R. LYU, M. (2008). Semi-supervised svm batch mode active learning for image retrieval. In *IEEE conference on Computer Vision and Pattern Recognition*. 61, 62, 66, 68, 70, 71

HUANG, C.H., YEH, Y.R. & WANG, Y.C.F. (2012a). Recognizing actions across cameras by exploring the correlated subspace. In *European Conference on Computer Vision*. 47

HUANG, D., CABRAL, R.S. & DE LA TORRE, F. (2012b). Robust regression. In *European Conference on Computer Vision*, 616–630, Springer. 29

HUISKES, M.J. & LEW, M.S. (2008). The mir flickr retrieval evaluation. In *ACM International Conference on Multimedia Information Retrieval*. 61, 62, 66, 68, 70, 71

JACOB, L., BACH, F. & VERT, J. (2008). Clustered multi-task learning: A convex formulation. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 81

JALALI, A., RAVIKUMAR, P., SANGHAVI, S. & RUAN, C. (2010). A dirty model for multi-task learning. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 9, 18, 29, 81, 95, 101

JI, S. & YE, J. (2009). An accelerated gradient method for trace norm minimization. In *International Conference on Machine Learning*. 81

JIANG, L., HAUPTMANN, A.G. & XIANG, G. (2012). Leveraging high-level and low-level features for multimedia event detection. In *ACM International Conference on Multimedia*. 78

JIANG, Z., LIN, Z. & DAVIS, L.S. (2011). Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *IEEE Conference on Computer Vision and Pattern Recognition*. 80

JIE, L., TOMMASI, T. & CAPUTO, B. (2011). Multiclass transfer learning from unconstrained priors. In *IEEE International conference on Computer Vision*. 95, 101

JUNEJO, I.N., DEXTER, E., LAPTEV, I. & PEREZ, P. (2011). View-independent action recognition from temporal self-similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**. 24, 25, 28, 31, 32, 44, 45, 46

KANG, Z., GRAUMAN, K. & SHA, F. (2011). Learning with whom to share in multi-task feature learning. In *International Conference on Machine Learning*. 9

KRISHNAN, D. & FERGUS, R. (2009). Fast image deconvolution using hyper-laplacian priors. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 91

LAN, Z., BAO, L., YU, S.I., LIU, W. & HAUPTMANN, A.G. (2012). Double fusion for multimedia event detection. In *International Conference on MultiMedia Modeling*. 79, 94

LANZ, O. (2006). Approximate bayesian multibody tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**, 1436–1449. 12, 17, 22

LAPTEV, I. (2005). On space-time interest points. *International Journal of Computer Vision*, **64**, 107–123. 74, 79

LAPTEV, I., MARSZALEK, M., SCHMID, C. & ROZENFELD, B. (2008). Learning realistic human actions from movies. In *IEEE conference on Computer Vision and Pattern Recognition*. 24, 42

LEE, H., BATTLE, A., RAINA, R. & NG, A.Y. (2006). Efficient sparse coding algorithms. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 80

LEPRI, B., SUBRAMANIAN, R., KALIMERI, K., STAIANO, J., PIANESI, F. & SEBE, N. (2012). Connecting meeting behavior with extraversion: A systematic study. *IEEE Trans. on Affective Computing*, **3**, 443–455. 6

LI, B., CAMPS, O.I. & SZNAIER, M. (2012). Cross-view activity recognition using hankelets. In *IEEE conference on Computer Vision and Pattern Recognition*. 25, 47, 48

LI, R. & ZICKLER, T. (2012). Discriminative virtual views for cross-view action recognition. In *IEEE conference on Computer Vision and Pattern Recognition*. 25, 47

LIN, D. (1998). An information-theoretic definition of similarity. In *International Conference on Machine Learning*. 83

LIN, Z., JIANG, Z. & DAVIS, L. (2010). Recognizing actions by shape-motion prototype trees. In *IEEE International conference on Computer Vision*. 24

LIU, H. & YU, L. (2003). Feature selection for high - dimensional data: A fast correlation-based filter solution. In *International conference on Machine Learning*. 64

LIU, J., LUO, J. & SHAH, M. (2009). Recognizing realistic actions from videos in the wild. In *IEEE conference on Computer Vision and Pattern Recognition*. 61, 62, 68, 70, 71

LIU, J., SHAH, M., KUIPERS, B. & SAVARESE, S. (2011). Cross-view action recognition via view knowledge transfer. In *IEEE conference on Computer Vision and Pattern Recognition*. 25, 28, 47

LOWE, D.G. (2004). Distinctive image features from scale invariant key points. *International Journal of Computer Vision*, **60**, 91–110. 74, 79

LUO, J., YU, J., JOSHI, D. & HAO, W. (2008). Event recognition: viewing the world with a third eye. In *ACM International Conference on Multimedia*. 74

LUO, Y., TAO, D., GENG, B., XU, C. & MAYBANK, S. (2013). Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE Transactions on Image Processing*, **22**, 523–536. 29

MA, Z., NIE, F., YANG, Y., UIJLINGS, J. & SEBE, N. (2012). Web image annotation via subspace-sparsity collaborated feature selection. *IEEE Transactions on Multimedia*, **14**, 1021–1030. 56

MAHASSENI, B. & TODOROVIC, S. (2013). Latent multitask learning for view-invariant action recognition. In *International Conference on Computer Vision*. 26, 30, 47, 48

MAHMOOD, T., VASILESCU, A. & SETHI, S. (2001). Recognizing action events from multiple viewpoints. In *IEEE International conference on Computer Vision*. 24

MAIRAL, J., BACH, F., PONCE, J., SAPIRO, G. & ZISSERMAN, A. (2008). Discriminative learned dictionaries for local image analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*. 79, 87

MAIRAL, J., BACH, F., PONCE, J. & SAPIRO, G. (2009a). Online dictionary learning for sparse coding. In *International Conference on Machine Learning*. 80, 88, 90, 91

MAIRAL, J., BACH, F., PONCE, J., SAPIRO, G. & ZISSERMAN, A. (2009b). Non-local sparse models for image restoration. In *IEEE International conference on Computer Vision*. 79

MAURER, A., PONTIL, M. & PAREDES, B.R. (2013). Sparse coding for multitask and transfer learning. In *International Conference on Machine Learning*. 81

MOXLEY, E., MEI, T. & MANJUNATH, B. (2010). Video annotation through search and graph reinforcement mining. *IEEE Transactions on Multimedia*, **12**, 184–193. 51

MUÑOZ-SALINAS, R., YEGUAS-BOLIVAR, E., SAFFIOTTI, A. & CARNICER, R.M. (2012). Multi-camera head pose estimation. *Mach. Vis. Appl.*, **23**, 479–490. 6, 9, 17, 18, 19

MURPHY-CHUTORIAN, E. & TRIVEDI, M.M. (2009). Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 607–626. 6

NATARAJAN, P., WU, S., VITALADEVUNI, S., ZHUANG, X., TSAKALIDIS, S., PARK, U., PRASAD, R. & NATARAJAN, P. (2012). Multimodal feature fusion for robust event detection in web videos. In *IEEE Conference on Computer Vision and Pattern Recognition*. 78, 94

NENE, S.A., NAYAR, S.K. & MURASE, H. (1996). Columbia object image library (coil-20). In *Technical Report CUCS-005-96*. 60, 62, 68, 70, 71

NI, B., WANG, G. & MOULIN, P. (2011). Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *International Conference on Computed Vision Workshops*, 1147–1153. 43

NIE, F., HUANG, H., CAI, X. & DING, C. (2010). Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 53, 62, 66

NIST (2013). Nist trecvid multimedia event detection. `http://www.nist.gov/itl/iad/mig/med13.cfm`. 74, 82, 93

OBOZINSKI, G., TASKAR, B. & JORDAN, M.I. (2007). Multi-task feature selection. In *UC Berkeley Technical Report*. 56

OROZCO, J., GONG, S. & XIANG, T. (2009). Head pose classification in crowded scenes. In *British Machine Vision Conference*. 6, 8

PARAMESWARAN, V. & CHELLAPP, R. (2005). Human action-recognition using mutual invariants. In *Computer Vision and Image Understanding*. 28

PENG, H., LONG, F. & DING, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226–1238. 51, 64

POPPE, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, **28**, 976–990. 24

QUATTONI, A., CARRERAS, X., COLLINS, M. & DARRELL, T. (2009). An efficient projection for $l_{1,\infty}$ regularization. In *International Conference on Machine Learning*. 81

RAJAGOPAL, A., SUBRAMANIAN, R., VIERIU, R., RICCI, E., LANZ, O., SEBE, N. & RAMAKRISHNAN, K. (2012). An adaptation framework for head pose estimation in dynamic multi-view scenarios. In *Asian conference on Computer Vision*. 6, 9, 17, 18, 19

RAO, C., YILMAZ, A. & SHAH, M. (2002). View-invariant representation and recognition of actions. *International Journal of Computer Vision*, **50**, 203–226. 25, 28

REDDY, K., LIU, J. & SHAH, M. (2009). Incremental action recognition using feature tree. In *IEEE International conference on Computer Vision*. 47

RESNIK, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *International Conference on Artificial Intelligence*. 84

SABERIAN, M.J., MASNADI-SHIRAZI, H. & VASCONCELOS, N. (2011). Taylorboost: First and second-order boosting algorithms with explicit margin control. In *IEEE conference on Computer Vision and Pattern Recognition*. 57

SADLIER, D.A. & O'CONNOR, N.E. (2005). Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Transactions on Circuits and System for Video Technology*, **15**, 1225–1233. 78

SCHLKOPF, B., SMOLA, A.J. & MLLER, K.R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**, 1299–1319. 57

SHE, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics*, **3**, 384–415. 91

SIGAL, L., BALAN, A. & BLACK, M.J. (2010). HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal on Computer Vision*, **87**, 1–2. 61, 62, 66, 68, 70, 71

SIN (2013). Nist trecvid semantic indexing. http://www-nlpir.nist.gov/projects/tv2013/tv2013.html#sin. 75, 82, 85, 92

SMEATON, A., OVER, P. & KRAAIJ, W. (2006). Evaluation campaigns and trecvid. In *ACM Multimedia Information Retrieval*. 79

SNOEK, C. & SMEULDERS, A. (2010). Visual-concept search solved? *IEEE Computer*, **43**, 76–78. 74, 75, 79, 84

SNOEK, C., WORRING, M., VAN GEMERT, J., GEUSEBROEK, J. & SMEULDERS, A. (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM International Conference on Multimedia*. 74

STIEFELHAGEN, R., BOWERS, R. & JONATHAN, G.F. (2007). Multimodal technologies for perception of humans, CLEAR. 17

STRUBE, M. & PONZETTO, S.P. (2006). Wikirelate! computing semantic relatedness using wikipedia. In *Conference on Artificial Intelligence (AAAI)*. 82

TAMRAKAR, A., ALI, S., YU, Q., LIU, J., JAVED, O., DIVAKARAN, A., CHENG, H. & SAWHNEY, H. (2012). Evaluation of low-level features and their combinations for complex event detection in open source videos. In *IEEE Conference on Computer Vision and Pattern Recognition*. 78

TANG, K., KOLLER, D. & LI, F.F. (2012). Learning latent temporal structure for complex event detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 78

TOSATO, D., FARENZENA, M., CRISTANI, M., SPERA, M. & MURINO, V. (2010). Multi-class classification on riemannian manifolds for video surveillance. In *European conference on Computer Vision*. 6, 8, 18, 19

TSAI, M.H., WANG, J., ZHANG, T., GONG, Y. & HUANG, T.S. (2011). Learning semantic embedding at a large scale. In *IEEE International Conference on Image Processing*, 2497–2500. 29

VAHDAT, A., CANNONS, K., MORI, G., OH, S. & KIM, I. (2013). Compositional models for video event detection: A multiple kernel learning latent variable approach. In *IEEE International conference on Computer Vision*. 78

VOIT, M. & STIEFELHAGEN, R. (2009). A system for probabilistic joint 3d head tracking and pose estimation in low-resolution, multi-view environments. In *Computer Vision Systems*, 415–424. 6, 9

WAGNER, A., WRIGHT, J., GANESH, A., ZHOU, Z., MOBAHI, H. & MA, Y. (2012). Towards a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**. 51

WANG, G., CHUA, T.S. & ZHAO, M. (2008). Exploring knowledge of subdomain in a multi-resolution bootstrapping framework for concept detection in news video. In *ACM International Conference on Multimedia*. 78

WANG, H., KLÄSER, A., SCHMID, C. & LIU, C.L. (2011). Action recognition by dense trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition*. 79

WANG, Y., MEI, T., GONG, S. & HUA, X.S. (2009). Combining global, regional and contextual features for automatic image annotation. *Pattern Recognition*, **42**, 259–266. 51

WEINLAND, D., BOYER, E. & RONFARD, R. (2007). Action recognition from arbitrary views using 3d exemplars. In *IEEE International conference on Computer Vision*. 24, 25, 27, 39, 42, 47

WEINLAND, D., ÖZUYSAL, M. & FUA, P. (2010). Making action recognition robust to occlusions and viewpoint changes. In *European Conference on Computer Vision*. 24, 25, 27, 39, 42, 48

WEINLAND, D., RONFARD, R. & BOYER, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, **115**, 224–241. 24, 74

WRIGHT, J., YANG, A.Y., SASTRY, S. & MA, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 210–227. 53

WU, X. & JIA, Y. (2012). View-invariant action recognition using latent kernelized structural SVM. In *European Conference on Computer Vision*. 47, 48

XU, C., WANG, J., WAN, K., LI, Y. & DUAN, L. (2006). Live sports event detection based on broadcast video and web-casting text. In *ACM International Conference on Multimedia*. 78

YAN, P., KHAN, S.M. & SHAH, M. (2008). Learning 4d action feature models for arbitrary view action recognition. In *IEEE conference on Computer Vision and Pattern Recognition*. 27

YAN, Y., SUBRAMANIAN, R., LANZ, O. & SEBE, N. (2012). Active transfer learning for multi-view head-pose classification. In *International conference on Pattern Recognition*. 6

YAN, Y., LIU, G., RICCI, E. & SEBE, N. (2013a). Multi-task linear discriminant analysis for multi-view action recognition. In *International Conference on Image Processing*, 2842–2846. 9, 30, 53

YAN, Y., RICCI, E., SUBRAMANIAN, R., LANZ, O. & SEBE, N. (2013b). No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *International Conference on Computer Vision*. 29, 51

YAN, Y., XU, Z., LIU, G., MA, Z. & SEBE, N. (2013c). Glocal structural feature selection with sparsity for multimedia data understanding. In *ACM International Conference on Multimedia*. 54, 56

YANAGAWA, A., C. LOUI, A., LUO, J., CHANG, S.F., ELLIS, D., JIANG, W., KENNEDY, L. & LEE, K. (2008). Kodak consumer video benckmark data set: concept definition and annotation. In *Columbia University ADVENT Technical Report 246-2008-4, Sep*. 62, 66, 68, 70, 71

YANG, J., YAN, R. & HAUPTMANN, A. (2007). Cross-domain video concept detection using adaptive svms. In *ACM International Conference on Multimedia*. 57

YANG, Y., XU, D., NIE, F., LUO, J. & ZHUANG, Y. (2009). Ranking with local regression and global alignment for cross media retrieval. In *ACM International Conference on Multimedia*. 55, 79

YANG, Y., SHEN, H., MA, Z., HUANG, Z. & XIAOFANG, Z. (2011). $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *International Joint Conference on Artificial Intelligence*. 53, 64, 66

YANG, Y., NIE, F., XU, D., LUO, J., ZHUANG, Y. & PAN, Y. (2012). A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**, 723–742. 53

YE, J. (2007). Least squares linear discriminant analysis. In *International Conference on Machine Learning*. 25, 34

YILMAZ, A. & SHAH, M. (2005). Actions sketch: A novel action representation. In *IEEE conference on Computer Vision and Pattern Recognition*. 24

YU, S., XU, Z., DING, D., SZE, W., VICENTE, F., LAN, Z., CAI, Y., RAWAT, S., SCHULAM, P., MARKANDAIAH, N., BAHMANI, S., JUAREZ, A., TONG, W., YANG, Y., BURGER, S., METZE, F., SINGH, R., RAJ, B., STERN, R., MITAMURA, T., NYBERG, E. & HAUPTMANN, A.G. (2012). Informedia e-lamp @ trecvid2012: Multimedia event detection and recounting med and mer. In *NIST TRECVID workshop*. 78, 94

YUAN, X.T. & YAN, S. (2010). Visual classification with multi-task joint sparse representation. In *IEEE conference on Computer Vision and Pattern Recognition*. 51, 81

ZABULIS, X., SARMIS, T. & ARGYROS, A. (2009). 3d headpose estimation from multiple distant views. In *British Machine Vision Conference*. 6, 9, 17, 19

ZHANG, M.L. & ZHOU, Z.H. (2007). Ml-kNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, **40**, 2038–2048. 64

ZHANG, Q. & LI, B. (2010). Discriminative K-SVD for dictionary learning in face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 87

ZHANG, T., GHANEM, B., LIU, S. & AHUJA, N. (2012). Robust visual tracking via multi-task sparse learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 29, 81

ZHANG, Y. & YEUNG, D. (2010). A convex formulation for learning task relationships in multi-task learning. In *Proceedings of Conference on Uncertainty in Artificial Intelligence*. 81

ZHANG, Y. & YEUNG, D.Y. (2011). Multi-task learning in heterogeneous feature spaces. In *Conference on Artificial Intelligence (AAAI)*. 29

ZHANG, Z. & ZHA, H. (2002). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. In *SIAM Journal of Scientific Computing*. 55

ZHONG, L.W. & KWOK., J.T. (2012). Convex multitask learning with flexible task clusters. In *International Conference on Machine Learning*. 9, 18, 20, 29

ZHOU, J., CHEN, J. & YE, J. (2011a). Clustered multi-task learning via alternating structure optimization. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*. 9, 18, 20, 29, 81

ZHOU, J., CHEN, J. & YE, J. (2011b). *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University. 21

ZHOU, Q., WANG, G., JIA, K. & ZHAO, Q. (2013). Learning to share latent tasks for action recognition. In *IEEE International conference on Computer Vision*. 29

ZHU, X., GHAHRAMANI, Z. & LAFFERTY, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *International conference on Machine Learning*. 57

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67**, 301–320. 84

ZUO, W., MENG, D., ZHANG, L., FENG, X. & ZHANG, D. (2013). A generalized iterated shrinkage algorithm for non-convex sparse coding. In *IEEE International conference on Computer Vision*. 91