Recognizing Daily Activities from First-person Videos with Multi-task Clustering

Yan Yan[1], Elisa Ricci[2,3], Gaowen Liu[1], Nicu Sebe[1]

[1] *Dept. of Information Engineering and Computer Science, University of Trento, Italy*
[2] *Fondazione Bruno Kessler, Italy*
[3] *Dept. of Engineering, University of Perugia, Italy*

Abstract. The widespread adoption of low-cost wearable devices requires novel paradigms for analysing human behaviour. In particular, when focusing on first-person cameras continuously recording several hours of the users life, the task of activity recognition is especially challenging. As a huge amount of unlabeled data is automatically generated in this scenario, despite recent notable attempts, more scalable algorithms and more effective feature representations are required. In this paper, we address the problem of *everyday activity recognition* from visual data gathered from a *wearable camera* proposing a novel *multi-task learning* framework. We argue that, even if label information is not provided, we can take advantage of the fact that the tasks of recognizing activities of daily life of multiple individuals are related, *i.e.* typically people tend to perform the same actions in the same environment (*e.g.* people at home in the morning typically have breakfast and brush their teeth). To exploit this information we propose a novel multi-task clustering approach. With our method, rather than clustering data from different users separately, we look for data partitions which are similar among related tasks. Thorough experiments on two publicly available first-person vision datasets demonstrate that the proposed approach consistently and significantly outperforms several state-of-the-art methods.

1 Introduction

Human behaviour analysis is an important research area in computer vision. Automatically understanding *what people do* by analyzing visual streams recorded from surveillance cameras is a challenging task and implies recognizing the activities of the people of interest, the environment where they operate, the other people with whom they interact, the objects they manipulate and even their future intentions. While many progresses have been made in this area, recent works [1] have demonstrated as the traditional "third-person" view perspective (*i.e.* employing fixed cameras mounted all around in the user's environment) may be insufficient for understanding people activities and intentions and that wearable cameras can provide an alternative or complementary source of information. Wearable cameras can be employed in many different applications, such as in driver's assistance systems, for monitoring assembly operations in manufacturing, in ambient assisted living and, more recently, in the context of the so

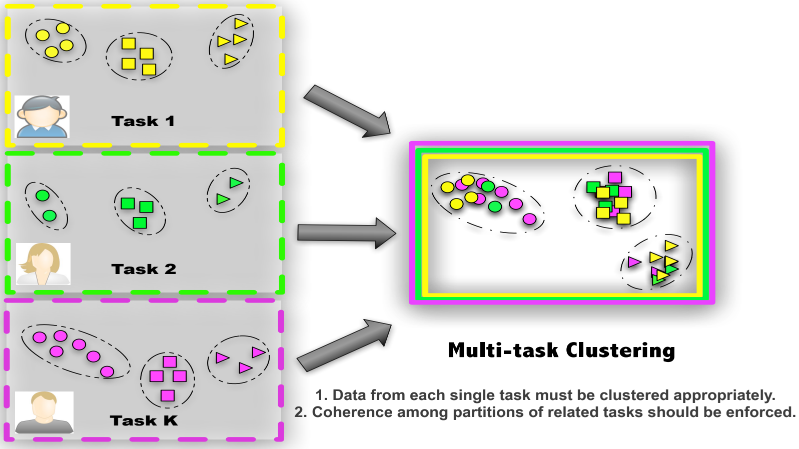

Task 1
Task 2
Task K
Multi-task Clustering
1. Data from each single task must be clustered appropriately.
2. Coherence among partitions of related tasks should be enforced.

Fig. 1. Overview of our proposed multi-task clustering approach for First Person Vision activity recognition (Figure is best viewed in color and zoom).

called "life-logging" [2,3] (*i.e.* where a first-person camera continuously records a whole day of its wearer life).

In this paper, we focus specifically on everyday activity recognition from a "first-person" vision (FPV) perspective. This problem poses several challenges. With wearable cameras typically several hours of videos are recorded. This generates a large amount of data for which labels are not available as the annotation would require a massive human effort. Thus, for accurate recognition, algorithms which are both scalable and able to operate in an unsupervised setting are required. Moreover, designing effective visual features representations in this unconstrained FPV scenario is much more challenging than in the case of fixed cameras. In this paper, we propose to address the problem of everyday activity recognition from unlabeled visual data within a multi-task learning framework. When considering the tasks of recognizing activities of daily living of many individuals, it is natural to assume that these tasks are related. For example, people working in an office environment tend to perform the same kind of activities (*e.g.* typing on keyboard in front of a personal computer, reading and writing documents). Similarly, most people when they wake up in the morning use to drink coffee and brush their teeth. Thus, it is intuitive that, when performing activity recognition, learning from data of all the individuals simultaneously is advantageous with respect to considering each person separately. However, the data distributions of single tasks can be different, since visual data corresponding to different people may exhibit different features. In particular if there are limited data for a single person, typical clustering methods may fail to discover the correct clusters. In this case, using data from other people as an auxiliary source of information can improve clustering results. However, simply combining data from different people together and applying traditional clustering approach does not necessarily increase accuracy, because the data distributions of single tasks can be different, violating *i.i.d.* assumptions. To address this problem, we propose to invoke the novel paradigm of multi-task clustering (MTC). Specifically, we introduce two novel methods, derived by a common framework based on the minimization of an objective function balancing two terms, one which ensures the data of each single task to be clustered appropriately, the other which enforces some coherence between the clustering results of related tasks. We demonstrate the effectiveness of our approaches on two recent FPV datasets, the FPV activity of daily living dataset [3] and the coupled ego-motion and eye-motion dataset introduced in [4], comparing them with several single task and multi-task learning methods. Fig. 1 depicts an overview of the proposed method.

The main contributions of this work are the following: (i) To our knowledge, this paper is the first to address the problem of everyday activity recognition within a MTC framework. While our framework can be used to analyze visual streams recorded from fixed cameras, we tackle the more challenging scenario of egocentric vision. (ii) The two proposed multi-task clustering approaches are novel and two efficient algorithms are derived for solving the associated optimization problems. (iii) Our experimental evaluation demonstrates that, independently of the adopted feature representations, a multi-task learning framework is greatly advantageous for FPV activity recognition with respect to traditional single task approaches. (iv) The proposed MTC algorithms are general and can be applied to many other computer vision and pattern recognition problems.

2 Related Works

Activity Recognition in Egocentric Videos. Analysing human behaviors from data collected from wearable devices has received considerable attention recently, not only in computer vision but also in other related research areas, *e.g.* ubiquitous computing [5, 6]. While many recent works are based on the use of RFID tags or inertial sensors, systems based on first-person cameras still play an important role being generally cheap and easy to deploy. Aghazadeh *et al.* [7] considered the problem of discovering anomalous events analysing the video stream captured from a small video camera attached to a person's chest. In [2] a summarization approach targeted to egocentric videos is presented. Fathi *et al.* [8] introduced a method for individuating social interactions in first-person videos collected during social events. Some recent works have faced the multiple challenges of recognizing complex activities of everyday life from an egocentric perspective in different scenarios (*e.g.* kitchen, office, home) [3, 4, 9, 10]. In [3] the authors demonstrated the importance of using features based on object detectors' output in the challenging unconstrained scenario of everyday at home activity recognition. In [9] RGB-D sensors are employed for fine-grained recognition of kitchen activities. In [4] the task of recognizing egocentric activities in an office environment is considered and motion descriptors extracted from an outside looking camera are used jointly with features modeling the eye movements of the wearer captured by an inside looking camera. In [10] activity recognition in a kitchen scenario (multiple subjects preparing different recipes) is considered.

A codebook learning framework is proposed in order to alleviate the problem of the large within-class data variability due to the different execution styles and speed among different subjects.

In this paper, we address the problem of analysing activities of daily living under the perspective of multi-task learning. Multi-task learning methods have been previously investigated in the context of visual-based activity recognition from fixed cameras and in a supervised setting [11–13]. In this paper, we consider the more challenging scenario where no annotated data are provided, which is typical when analyzing visual streams from wearable cameras.

Multi-task Learning. Recently multi-task learning (MTL) approaches [14] have demonstrated their effectiveness in several applications in computer vision, such as object detection [15], indoor localization [16], face verification [17] or head pose estimation [18]. The idea of multi-task learning is simple: learning from data of multiple related tasks simultaneously produces more accurate classification and regression models with respect to learning on every single task independently. While many works have introduced supervised MTL approaches, only few have considered an unsupervised setting [19–21], *i.e.* the scenario where all the data are unlabeled and the aim is to predict the cluster labels in each single task. Gu *et al.* [19] proposed to learn a subspace shared by all the tasks, through which the knowledge of one task can be transferred to all the others. Zhang *et al.* [21] introduced a MTC approach based on a pairwise agreement term which encourages coherence among clustering results of multiple tasks. In [20] the k-means algorithm is revised from a Bayesian nonparametric viewpoint and extended to MTL.

In this paper, we propose two novel approaches for multi-task clustering. The first one is inspired by the work in [21] but it is based on another objective function and thus on a radically different optimization algorithm. Furthermore, in the considered application, it provides superior accuracy with respect to [21]. Our second approach instead permits to easily integrate prior knowledge about the tasks and the data of each task (*e.g.* temporal consistency among subsequent video clips). Moreover, it relies on a convex optimization problem, thus avoids the issues related to local minima of previous methods [19–21].

3 Multi-task Clustering for FPV Activity Recognition

In this paper, we focus on the problem of everyday activity recognition from wearable cameras. More specifically, we consider several video clips collected by a certain number of people involved in activities of daily living. No labeled data are provided. We only assume that people perform about the same tasks, a very reasonable assumption in the context of everyday activity analysis. Considering each individual's data as a specific task, we propose a MTC approach. To stress the generality of our method, we apply it in two different scenarios: an office environment where people are involved in typical activities such as browsing the web or writing documents and a home environment where a chest mounted camera records users' activities such as opening a fridge or preparing tea. To

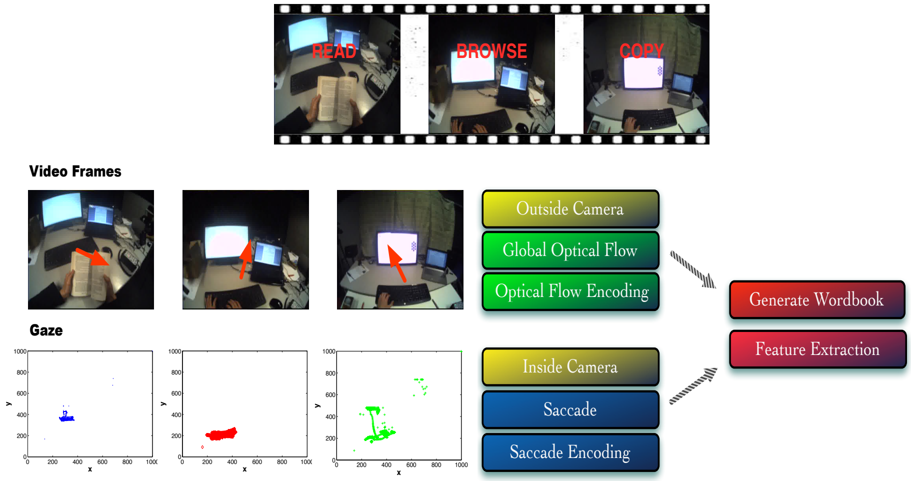


Fig. 2. Feature extraction pipeline on the FPV office dataset. Some frames corresponding to the actions *read*, *browse* and *copy* are shown together with the corresponding optical flow features (top) and eye-gaze patterns depicted on the 2-D plane (bottom). It is interesting to observe the different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

perform experiments we use two publicly available datasets, corresponding to the scenarios described above: the FPV office dataset introduced in [4] and the FPV activity of daily living dataset described in [3]. Both datasets contains visual streams recorded from an outside-looking wearable camera while the office dataset also has informations about eye movements acquired by an inside-looking camera. Further details about the datasets are provided in the experimental section. In the following we describe the adopted feature descriptors and the proposed MTC framework.

3.1 Features Extraction in Egocentric Videos

Due to the large variability of visual data collected from wearable cameras there exist no typical feature descriptors but different representations are adopted dependently on the context. While in some situations extracting motion information by computing optical flow vectors may suffice [4], in other cases motion patterns may be too noisy and other kind of informations (*e.g.* presence/absence of objects) must be exploited. In this paper we demonstrate that, independently from the employed feature descriptors, MTC is an effective strategy for recognizing everyday activities. We now describe the adopted feature representations respectively for the considered office and home scenarios.

FPV office dataset. We follow [4] and extract features describing both the eye motion (obtained by the inside-looking camera) and the head and body motion (computed processing the outside camera's stream). To calculate the eye motion features, we consider the gaze coordinates provided in the dataset and smooth them applying a median filter. Then the continuous wavelet transform is adopted for saccade detection separately on the x and y motion components

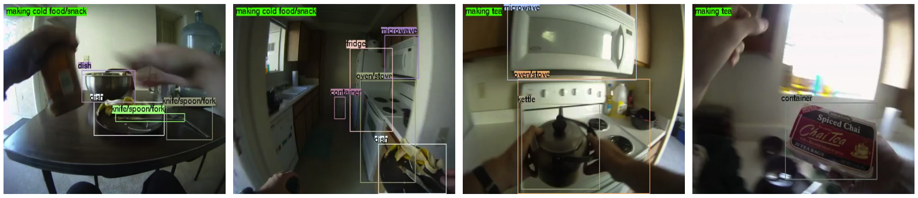


Fig. 3. FPV home dataset: frames depicting examples of the activities *making cold food/snack* and *making tea* and the detected objects.

[22]. The resulting signals are quantized according to magnitude and direction and are coded with a sequence of discrete symbols. To analyze the streams of the output camera, for each frame the global optical flow is computed by tracking corner points over consecutive frames and taking the mean flow in the x and y directions. Then, the optical flow vectors are quantized according to magnitude and direction with the same procedure adopted in the eye motion case. The obtained sequences of symbols are then processed to get the final video clip descriptors. We use a temporal sliding window approach to build a n-gram dictionary over all the dataset. Then each video is divided into segments corresponding to 15 seconds, each of them representing a video clip. For each sequence of symbols associated to a video clip, a histogram over the dictionary is computed. The final feature descriptor $\mathbf{x}_i$ is calculated by considering some statistics over the clip histogram and specifically computing the maximum, the average, the variance, the number of unique n-grams, and the difference between maximum and minimum count. Fig.2 shows the feature extraction pipeline.

FPV home dataset. We adopt the same object-centric approach proposed in [3], *i.e.* to compute features for each video clip, we consider the output of several object detectors. More specifically, we use the pre-segmented video clips and the active object models in [3]. Active object models are introduced to exploit the fact that objects look different when being interacted with (*e.g.* open and close fridge). Therefore in [3] additional detectors are trained using a subset of training images depicting object appearance when objects are used by people. Fig.3 shows some frames associated to the activities *making cold food/snack* and *making tea*: the output of the object detectors are depicted. To obtain object-centric features for each frame a score for each object model and each location is computed. Then the maximum scores of all the object models are used as frame features. To compute the final clip descriptors $\mathbf{x}_i$, two approaches are adopted: one based on "bag of features" (accumulating frame features over time) and the other based on temporal pyramids. The temporal pyramid features are obtained concatenating several histograms constructed with accumulation: the first is a histogram over the full temporal extent of a video clip, the next is the concatenation of two histograms obtained by temporally segmenting the video into two parts, and so on.

3.2 Multi-task Clustering

We consider T related tasks corresponding to T different people. For each task t, a set of data samples $X^t = \{\mathbf{x}_1^t, \mathbf{x}_2^t, ..., \mathbf{x}_{N_t}^t\}$ is available, where $\mathbf{x}_j^t \in I\!R^d$ is the

d-dimensional feature vector describing the j-th video clip and N_t is the total number of samples associated to the t-th task. In the following we denote with $(\cdot)'$ the transpose operator, $N = \sum_{t=1}^{T} N_t$ is the total number of datapoints, while $\mathbf{X} \in \mathbb{R}^{N \times d}$, $\mathbf{X} = [\mathbf{X}^{1'}\ \mathbf{X}^{2'}\ \ldots\ \mathbf{X}^{T'}]'$ is the data matrix obtained by concatenating the task specific matrices $\mathbf{X}^t = [\mathbf{x}_1^t\ \mathbf{x}_2^t\ ...\ \mathbf{x}_{N_t}^t]' \in \mathbb{R}^{N_t \times d}$. To discover people activities, we want to segment the entire video clip into parts, *i.e.* we want the data in the set X^t to be grouped into K_t clusters, where the number of required partitions can be different in different tasks. This is reasonable in the context of everyday activity recognition where people perform about the same activities. Furthermore, as we assume the tasks to be related, we also require that the resulting partitions are consistent with each other. This can be obtained by defining the following optimization problem:

$$\min_{\mathbf{\Theta}_t} \sum_{t=1}^{T} \Lambda(\mathbf{X}^t, \mathbf{\Theta}^t) + \sum_{t=1}^{T} \sum_{s=t+1}^{T} R(\mathbf{\Theta}^t, \mathbf{\Theta}^s) \qquad (1)$$

The problem (1) corresponds to the general problem of multi-task clustering, where the term $\Lambda(\cdot)$ represents a reconstruction error which must be minimized by learning the optimal task-specific model parameters $\mathbf{\Theta}^t$ (*i.e.* typically the cluster centroids and the associated assignment matrix), while $R(\cdot)$ is an "agreement" term imposing that, since the multiple tasks are related, also the associated model parameters should be similar. Under this framework, in this paper we propose two different approaches for MTC which mainly differ for the definition of the "agreement term". In the following subsections we present them in detail.

Notation. In the following $\mathbf{A}_{i\cdot}$, $\mathbf{A}_{\cdot j}$ denote respectively the i-th row and the j-th column of the matrix $\mathbf{A}$.

3.3 Earth Mover's Distance Multi-task Clustering

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

$$\min_{\mathbf{C}^t, \mathbf{W}^t, f_{ij}^{st} \geq 0} \sum_{t=1}^{T} \left\| \mathbf{X}^t - \mathbf{W}^t \mathbf{C}^t \right\|_F^2 + \lambda \sum_{t=1}^{T} \sum_{s=t+1}^{T} \sum_{i=1}^{K_t} \sum_{j=1}^{K_s} f_{ij}^{st} (\mathbf{C}_{i.}^t - \mathbf{C}_{j.}^s)'(\mathbf{C}_{i.}^t - \mathbf{C}_{j.}^s) \qquad (2)$$

$$\text{s.t.} \quad \sum_{j=1}^{K_s} f_{ij}^{st} = \sum_{n=1}^{N_t} \mathbf{W}_{ni}^t \quad \forall t, i \qquad \sum_{i=1}^{K_t} f_{ij}^{st} = \sum_{n=1}^{N_s} \mathbf{W}_{nj}^s \quad \forall s, j$$

$$\sum_{i=1}^{K_t} \sum_{j=1}^{K_s} f_{ij}^{st} = 1 \quad \forall s, t$$

The first term in the objective function is a relaxation of the traditional k-means objective function for T separated data sources. The second term, *i.e.* the agreement term, is added to explore the relationships between clusters of different data sources. It consists in the popular Earth Mover's Distance (EMD) [23] computed considering the signatures $\mathcal{T}$ and $\mathcal{S}$ obtained by clustering the data associated to task t and s separately, *i.e.* $\mathcal{T} = \{(\mathbf{C}^t_{1\cdot}, w^1_t),\ \dots,\ (\mathbf{C}^t_{K_t\cdot}, w^{K_t}_t)\}$, $w^i_t = \sum_{n=1}^{N_t} \mathbf{W}^t_{ni}$, and $\mathcal{S} = \{(\mathbf{C}^s_{1\cdot}, w^1_s),\ \dots,\ (\mathbf{C}^s_{K_s\cdot}, w^{K_s}_s)\}$, $w^i_s = \sum_{n=1}^{N_s} \mathbf{W}^s_{ni}$. In practice $\mathbf{C}^t_{i\cdot}$ and $\mathbf{C}^s_{j\cdot}$ are the cluster centroids and w^s_i, w^t_i denote the weights associated to each cluster (reflecting somehow the number of datapoints in each cluster). In practice the second term represents a sum of distances between two distributions and minimizing it we impose the found partitions between pairs of related tasks to be consistent. The variables f^{st}_{ij} are flow variables as follows from the definition of EMD as a transportation problem [23].

In (2) there are no constraints on the $\mathbf{C}_t$ values. In this paper we define the matrix $\mathbf{C} \in I\!R^{K\times d}$, $\mathbf{C} = [\mathbf{C}^{1\prime} \dots \mathbf{C}^{T\prime}]'$, $K = \sum_{t=1}^{T} K_t$, and we impose that the columns of $\mathbf{C}$ are a weighted sum of certain data points, *i.e.* $\mathbf{C} = \mathbf{PX}$ where $\mathbf{P} = \text{blkdiag}(\mathbf{P}^1, \dots, \mathbf{P}^T)$, $\mathbf{P} \in I\!R^{K\times N}$. In the following, for sake of simplicity and easy interpretation, we consider only a two tasks problem. The extension to T tasks is straightforward. Defining $\mathbf{F} = \text{diag}(f_{11} \dots f_{K_1K_2})$, $\mathbf{F} \in I\!R^{K_1K_2\times K_1K_2}$ and the block diagonal matrix $\mathbf{W} = \text{blkdiag}(\mathbf{W}^1, \mathbf{W}^2)$, $\mathbf{W} \in I\!R^{N\times K}$, the optimization problem (2) becomes:

$$\Delta(\mathbf{P}, \mathbf{W}, \mathbf{F}) = \min_{\mathbf{P},\mathbf{W},\mathbf{F}\geq 0} \|\mathbf{X} - \mathbf{WPX}\|_F^2 + \lambda \text{tr}(\mathbf{MPXX'P'M'F}) \tag{3}$$

$$\text{s.t.} \quad \|\mathbf{P}^t_{i\cdot}\|_1 = 1, \quad \forall i = 1, \dots, K \quad \forall\ t = 1, 2$$

$$\text{tr}(\mathbf{I}_j\mathbf{F}) = \sum_{i=1}^{N} \mathbf{W}_{ij}, \quad \forall j = 1, ..., K \tag{4}$$

$$\text{tr}(\mathbf{F}) = 1$$

where $\mathbf{I}_j \in I\!R^{K_1K_2\times K_1K_2}$ and $\mathbf{M} \in I\!R^{K_1K_2\times K}$ are appropriately defined selection matrices as

$$\mathbf{I}_j = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \mathbf{M} = \left[\begin{array}{cc} \underbrace{\begin{matrix} 1 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & \cdots \end{matrix}}_{1:K_1} & \underbrace{\begin{matrix} -1 & 0 & \cdots \\ 0 & -1 & \cdots \\ 0 & \cdots & -1 \\ -1 & 0 & \cdots \\ \vdots & \vdots & \vdots \\ 0 & \cdots & -1 \end{matrix}}_{K_1+1:K_1+K_2} \end{array} \right].$$

Optimization. To solve the proposed problem (3), we first note that the optimal solution of (3) can be found adopting an alternating optimization scheme, *i.e.* optimizing separately (3) first with respect to $\mathbf{P}$ and then with respect to $\mathbf{W}$ and $\mathbf{F}$ jointly. In both cases, a non-negative least square problem with constraints arises, for which standard solvers can be employed. However, due to computational efficiency, in this paper we consider an approximation of (3), replacing the constraints (4) with $\text{tr}(\mathbf{I}_j\mathbf{F}) = \mathbf{e}$, where $\mathbf{e} \in I\!R^{K_1K_2}$, $\mathbf{e}_i = \frac{1}{K_1}$, if $i \leq K_1$, $\mathbf{e}_i = \frac{1}{K_2}$ otherwise. This approximation implies that for each task the same number of datapoints is assigned to all the clusters. In this way a more efficient solver can be devised. Specifically, we adopt an alternating optimization strategy, *i.e.* we optimize (3) separately with respect to $\mathbf{F}$, $\mathbf{W}$ and $\mathbf{P}$ until convergence. The algorithm for solving (3) is summarized in Algorithm 1.

Kernelization. Finally, to kernelize the proposed method we consider a feature mapping $\phi(\cdot)$ and the associated kernel matrix $\mathbf{K}_\mathbf{X} = \phi(\mathbf{X})\phi(\mathbf{X})'$. The objective

```
Algorithm 1: Algorithm for solving (3).
Input: the data matrices $\mathbf{X}^1, \mathbf{X}^2$, the numbers of clusters $K_1$, $K_2$, the parameter $\lambda$.
Initialize $\mathbf{F}$ as an identity matrix.
Initialize $\mathbf{W} > 0$ with $l_1$ normalized columns and $\mathbf{P} > 0$ with $l_1$ normalized rows.
repeat
    Given $\mathbf{W}^k$, $\mathbf{P}^k$, compute $\mathbf{F}^{k+1}$ using a linear programming solver.
    Given $\mathbf{F}^{k+1}$, $\mathbf{P}^k$, compute $\mathbf{W}^{k+1}$ using a projected gradient method:
        $\mathbf{W}^{k+1} = \max(0, \mathbf{W}^k - \alpha_k \nabla_{\mathbf{W}} \Delta(\mathbf{P}^k, \mathbf{W}^k, \mathbf{F}^{k+1}))$.
    Given $\mathbf{F}^{k+1}$, $\mathbf{W}^{k+1}$, compute $\mathbf{P}^{k+1}$ using a projected gradient method:
        $\mathbf{P}^{k+1} = \max(0, \mathbf{P}^k - \alpha_k \nabla_{\mathbf{P}} \Delta(\mathbf{P}^k, \mathbf{W}^{k+1}, \mathbf{F}^{k+1}))$.
    Normalize $\mathbf{P}$ by $\mathbf{P}_{ij}^{k+1} \leftarrow \frac{\mathbf{P}_{ij}^{k+1}}{\sum_j \mathbf{P}_{ij}^{k+1}}$.
until convergence;
Output: the optimized matrices $\mathbf{W}, \mathbf{P}$.
```

function of (3) is:

$$\begin{gathered}\|\phi(\mathbf{X}) - \mathbf{WP}\ \phi(\mathbf{X})\|_F^2 + \lambda \mathrm{tr}(\mathbf{MP}\phi(\mathbf{X})\phi(\mathbf{X})'\mathbf{P}'\mathbf{M}'\mathbf{F}) = \\ \mathrm{tr}(\mathbf{K_X} - \mathbf{2K_X P'W'} + \mathbf{WPK_X P'W'} + \lambda \mathbf{MPK_X P'M'F})\end{gathered}$$

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

3.4 Convex Multi-task Clustering

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

$$\min_{\boldsymbol{\pi}_i^t} \sum_{t=1}^{T} \left(\sum_{i=1}^{N_t} \|\mathbf{x}_i^t - \boldsymbol{\pi}_i^t\|_2^2 + \lambda_t \sum_{\substack{i,j=1 \\ j>i}}^{N_t} w_{ij}^t \|\boldsymbol{\pi}_i^t - \boldsymbol{\pi}_j^t\|_1 \right) + \lambda_2 \sum_{\substack{t,s=1 \\ s>t}}^{T} \gamma_{st} \sum_{i=1}^{N_t} \sum_{j=1}^{N_s} \|\boldsymbol{\pi}_i^t - \boldsymbol{\pi}_j^s\|_2^2 \quad (5)$$

In (5) the first two terms guarantees that the data of each task are clustered: specifically with $\lambda_t = 0$ the found centroids are equal to the datapoints while as λ_t increases the number of different centroids $\boldsymbol{\pi}_i^t$ reduces. The last term in (5) instead imposes the found centroids to be similar if the tasks are related. The relatedness between tasks is modeled by the parameter γ_{st} which can be set using an appropriate measure between distributions. We consider the Maximum Mean Discrepancy $\mathcal{D}(X^t, X^s) = \| \frac{1}{N_t} \sum_{i=1}^{N_t} \phi(\mathbf{x}_i^t) - \frac{1}{N_s} \sum_{i=1}^{N_s} \phi(\mathbf{x}_i^s) \|^2$ [24], we computed it using a linear kernel and we set $\gamma_{st} = e^{-\beta \mathcal{D}(X^t, X^s)}$ with β being a user-defined parameter ($\beta = 0.1$ in our experiments). The parameters w_{ij}^t are used to enforce datapoints in the same task to be assigned to the same cluster and can be set according to some *a-priori* knowledge or in a way such that the found partitions structure reflects the density of the original data distributions.

Algorithm 2: Algorithm for solving (5).

```
Input: The data matrix $\mathbf{X}$, $\mathbf{E}, \mathbf{B}$, the parameter $\lambda_2$.
Set $\mathbf{Q} = \rho \mathbf{E}'\mathbf{E} + 2\mathbf{I} + 2\lambda_2 \mathbf{B}$.
Compute Cholesky factorization of the matrix $\mathbf{Q}$.
for $j$=1:$d$ do
    repeat
        Set $\mathbf{b}^k = \rho \mathbf{E}'\mathbf{q}^k - \mathbf{E}'\mathbf{p}^k + 2\mathbf{X}_{.j}$
        Update $\mathbf{\Pi}_{.j}$
        Solve $\mathbf{Q}[\mathbf{\Pi}_{.j}]^{k+1} = \mathbf{b}^k$
        Update $\mathbf{q}$ using the operator $ST_\lambda(x) = sign(x)\max(|x| - \lambda, 0)$
        $\mathbf{q}^{k+1} = ST_{1/\rho}(\mathbf{E}[\mathbf{\Pi}_{.j}]^{k+1} + \frac{1}{\rho}\mathbf{p}^k)$
        Update $\mathbf{p}$
        $\mathbf{p}^{k+1} = \mathbf{p}^k + \rho(\mathbf{E}[\mathbf{\Pi}_{.j}]^{k+1} - \mathbf{q}^{k+1})$
    until convergence;
Output: The final centroid matrix $\mathbf{\Pi}$.
```

Optimization. To solve (5) we propose an algorithm based on the alternating direction method of multipliers (ADMM) [25]. We consider the matrix $\mathbf{\Pi} = [\mathbf{\Pi}^{1'} \; \mathbf{\Pi}^{2'} \; \ldots \; \mathbf{\Pi}^{T'}]'$, $\mathbf{\Pi} \in \mathbb{R}^{N \times d}$, obtained concatenating the task-specific matrices $\mathbf{\Pi}^t = [\boldsymbol{\pi}_1^t \; \boldsymbol{\pi}_2^t \; ... \; \boldsymbol{\pi}_{N_t}^t]'$. The problem (5) can be solved considering d separate minimization subproblems (one for each column of $\mathbf{X}$) as follows:

$$\begin{aligned} &\min_{\mathbf{q},\, \mathbf{\Pi}_{.j}} \; \|\mathbf{X}_{.j} - \mathbf{\Pi}_{.j}\|_2^2 + \|\mathbf{q}\|_1 + \lambda_2 \|\mathbf{B}\mathbf{\Pi}_{.j}\|_2^2 \\ &\text{s.t.} \qquad \mathbf{E}\mathbf{\Pi}_{.j} - \mathbf{q} = 0 \end{aligned} \tag{6}$$

where $\mathbf{E}$ is a block diagonal matrix defined as $\mathbf{E} = \text{blkdiag}(\mathbf{E}^1, \mathbf{E}^2, \ldots, \mathbf{E}^T)$ and $\mathbf{E}^t \in \mathbb{R}^{|\mathcal{E}_t| \times N_t}$ is a matrix with $|\mathcal{E}_t| = \frac{N_t(N_t-1)}{2}$ rows. Each row is a vector of all zeros except in the position i where it has the value $\lambda_t w_{ij}^t$ and in the position j where it has the value $-\lambda_t w_{ij}^t$. Similarly the matrix $\mathbf{B} \in \mathbb{R}^{|\mathcal{B}| \times N}$, where $|\mathcal{B}| = \frac{T(T-1)}{2}$, imposes smoothness between the parameters of related tasks. A row of the matrix $\mathbf{B}$ is a vector with all zeros except in the terms corresponding to datapoints of the t-th task which are set to γ_{st} and to the terms corresponding to datapoints of the s-th task which are all set to $-\gamma_{st}$. To solve (6) we consider the associated lagrangian $L_\rho(\mathbf{\Pi}_{.j}, \mathbf{q}, \mathbf{p})$:

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

with $\mathbf{p}$ being the vector of augmented Lagrangian multipliers and ρ being the dual update step length. We devise an algorithm based on the ADMM where three steps, corresponding to the update of the three variables $\mathbf{\Pi}_{.j}, \mathbf{q}, \mathbf{p}$, are performed. We summarize our approach in Algorithm 2.

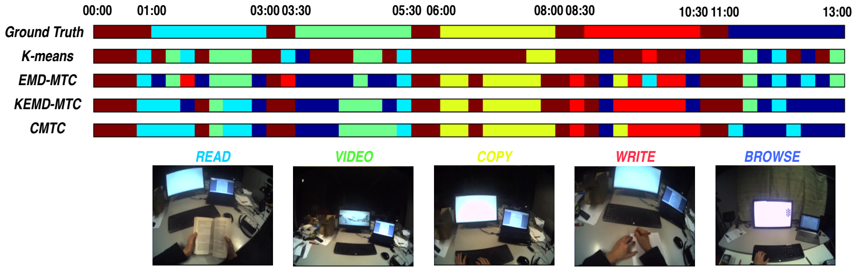


Fig. 4. FPV Office dataset. Temporal video segmentation on the second sequence of subject-3 (13 minutes): comparison of different methods. (Best viewed in color).

4 Experimental Results

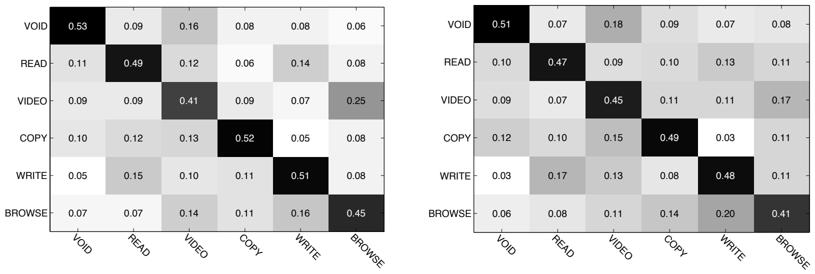
4.1 Datasets and Experimental Setup

The growing interest in the vision community towards novel approaches for FPV analysis has led to the creation of several publicly available datasets [2–4, 8]. In this paper we consider two of them, the FPV office dataset [4] and the FPV home dataset [3].

FPV office dataset [4]. This dataset consists of five activities which frequently occur in an office environment (*reading a book, watching a video, copying text from screen to screen, writing sentences on paper* and *browsing the internet*). Each action was performed by five subjects, who were instructed to execute each task for about two minutes, while 30 seconds intervals of void class were placed between target tasks. To provide a natural experimental setting, the void class contains a wide variety of actions such as conversing, singing and random head motions. The sequence of five actions was repeated twice to induce interclass variance. The dataset consists of over two hours of data, where the video from each subject is a continuous 25-30 minutes video.

FPV home dataset [3]. This dataset contains videos recorded from chest-mounted cameras by 20 different users. The users perform 18 non scripted daily activities in the house, like *brushing teeth, washing dishes,* or *making tea.* The length of the videos is in the range of 20-60 minutes. The annotations about the presence of 42 relevant objects (*e.g.* kettle, mugs, fridge) and about temporal segmentation are also provided.

Setup. In the experiments, we compare our methods (EMD Multi-task Clustering with linear and rbf kernel and Convex Multi-task Clustering denoted as EMD-MTC, KEMD-MTC, CMTC respectively) with single task clustering approaches, *i.e.* k-means (KM), kernel k-means (KKM), convex (CNMF) and semi (SemiNMF) nonnegative matrix factorization [26]. We also consider recent multi-task clustering algorithms such as the SemiEMD-MTC proposed in [21], its kernel version KSemiEMD-MTC and the LS-MTC method in [19]. To evaluate the clustering results, we adopt two metrics widely used in the literature: the clustering accuracy (Acc) and the normalized mutual information (NMI). For all the methods, except than for CMTC, 10 runs are performed corresponding to different initializations conditions. The average results are shown. In CMTC



	VOID	READ	VIDEO	COPY	WRITE	BROWSE
VOID	0.53	0.09	0.16	0.08	0.08	0.06
READ	0.11	0.49	0.12	0.06	0.14	0.08
VIDEO	0.09	0.09	0.41	0.09	0.07	0.25
COPY	0.10	0.12	0.13	0.52	0.05	0.08
WRITE	0.05	0.15	0.10	0.11	0.51	0.08
BROWSE	0.07	0.07	0.14	0.11	0.16	0.45

	VOID	READ	VIDEO	COPY	WRITE	BROWSE
VOID	0.51	0.07	0.18	0.09	0.07	0.08
READ	0.10	0.47	0.09	0.10	0.13	0.11
VIDEO	0.09	0.07	0.45	0.11	0.11	0.17
COPY	0.12	0.10	0.15	0.49	0.03	0.11
WRITE	0.03	0.17	0.13	0.08	0.48	0.11
BROWSE	0.06	0.08	0.11	0.14	0.20	0.41

Fig. 5. FPV Office dataset. Confusion matrices using saccade+motion features obtained with (left) KEMD-MTC and (right) CMTC methods.

the parameters λ_t are varied in order to obtain the desired number of clusters. The value of the regularization parameters of our approaches (λ for the methods based on EMD regularization and λ_2 for CMTC) are set in the range $\{0.01, 0.1, 1, 10, 100\}$.

4.2 Results

FPV office dataset [4]. To conduct experiments on this dataset, we consider $T = 5$ tasks, as the dataset contains videos corresponding to five people. As each datapoint corresponds to a video clip in this dataset, we set the parameters w_{ij}^t in CMTC in order to enforce temporal consistency, *i.e.* for each task t, $w_{ij}^t = 1$ if the features vectors $\mathbf{x}_i^t$ and $\mathbf{x}_j^t$ correspond to temporal adjacent video clips, otherwise $w_{ij}^t = 0$. Table 1 shows a comparison of the results associated to different clustering methods based on different types of features (*i.e.* only saccade, only motion and saccade+motion features). The last three rows correspond to methods which employ a non-linear kernel. From Table 1, several observations can be made. First, independently on the adopted features representation, multi-task clustering approaches always perform better than single task clustering methods (*e.g.* SemiEMD-MTC outperforms SemiNMF, EMD-MTC provide higher accuracy than CNMF, a value of λ_2 greater than 0 leads to an improvement in accuracy and NMI in CMTC). Confirming the findings reported in [4], we also observe that combining motion and saccade features is advantageous with respect to considering each single feature representation separately. Noticeably, our methods are among the best performers, with KEMD-MTC reaching the highest values of accuracy and NMI. This is somehow expected probably due to both the use of kernels and the adoption of the multi-task learning paradigm. Moreover, CMTC outperforms EMD-MTC by up to 4% which means that incorporating information about temporal consistency in the learning process is beneficial. Furthermore, in this case the use of Maximum Mean Discrepancy may capture better the relationship among tasks with respect to EMD. Fig.4 shows some qualitative temporal segmentation results on the second sequence of subject-3. In this case for example the CMTC methods outperforms all the other approaches and the effect of enforcing temporal consistency among clips is evident. More qualitative results are provided in the demo video in our supplementary material.

Table 1. Clustering results on FPV office dataset: comparison of different methods using saccade (S), motion (M) and saccade+motion (S+M) features.

	Avg Acc			Avg NMI		
	S	M	S+M	S	M	S+M
KM	0.230	0.216	0.257	0.029	0.021	0.045
SemiNMF [26]	0.320	0.303	0.358	0.149	0.131	0.166
SemiEMD-MTC [21]	0.371	0.349	0.415	0.229	0.209	0.259
LSMTC [19]	0.286	0.261	0.335	0.043	0.031	0.071
CNMF [26]	0.328	0.301	0.357	0.152	0.139	0.170
EMD-MTC	0.389	0.363	0.442	0.239	0.221	0.273
CMTC ($\lambda_2 = 0$)	0.367	0.346	0.413	0.224	0.209	0.244
CMTC	0.425	0.401	0.468	0.259	0.238	0.305
KKM	0.345	0.316	0.377	0.159	0.152	0.185
KSemiEMD-MTC [21]	0.387	0.359	0.432	0.241	0.228	0.287
KEMD-MTC	0.436	0.419	0.485	0.268	0.244	0.311

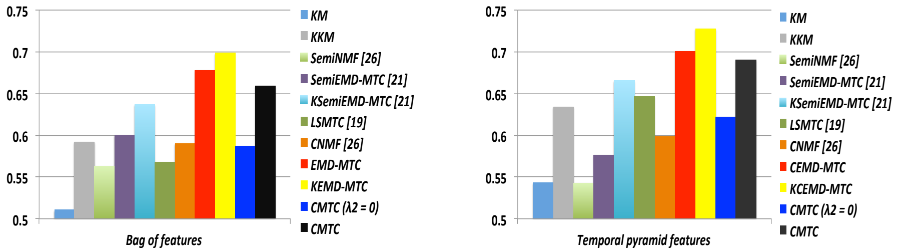


Fig. 6. Comparison of different methods using (left) bag of features and (right) temporal pyramid features on FPV home dataset. (Figure is best viewed in color).

Finally, Fig.5 shows the confusion matrices associated to our methods KEMD-MTC and CMTC. Examining the matrix associated to KEMD-MTC, we observe that the *void, copy* and *write* actions achieve relative high recognition accuracies compared with the *video* and *browse* actions. It is also interesting to note that 25% and 17% of the *video* actions are recognized as *browse* actions for KEMD-MTC and CMTC respectively, because of the similarity among motion and eye-gaze patterns.

FPV home dataset [3]. In this dataset there are 18 different non scripted activities. Since each person typically performs a small subset of the 18 activities, in our experiments we consider a series of three tasks problems, selecting videos associated to three randomly chosen users but imposing the condition that videos corresponding to the three users should have at least three activities in common. We perform 10 different runs. Fig.6 shows the results (average accuracy) obtained with different clustering methods for both the bag-of-words and the temporal pyramid features representation. In this series of experiments, we did not cluster video clips of fixed size as in the office dataset, but we consider the pre-segmented clips as provided with the dataset. In this scenario, it does not make sense to set w_{ij}^t in CMTC to model temporal consistency. Therefore, we set $w_{ij}^t = e^{-\|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2}$ if $e^{-\|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2} \leq \theta$ and $w_{ij}^t = 0$ otherwise. This is meant to enforce that the found partitions structure reflects the density of the original data distributions. Analyzing the results in Fig.6, it is evident that the MTC approaches outperforms their single task version (*e.g.* CMTC outperforms CMTC with $\lambda_2 = 0$, EMD-MTC outperforms CNMF, SemiEMD-MTC

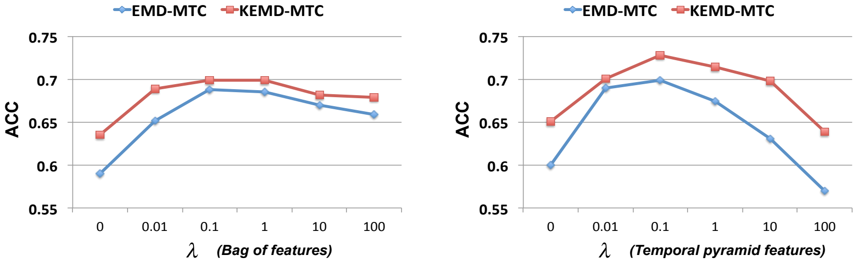


Fig. 7. FPV home dataset: performance variations of EMD-MTC and KEMD-MTC at different values of λ using (left) bag of features and (right) temporal pyramid features.

outperforms SemiNMF). On the other hand, our algorithms based on EMD regularization and CMTC achieve a considerably higher accuracy with respect to all the other methods. Finally, we also investigate the effect of different values of the regularization parameter λ in (3) on clustering performance. As shown in Fig.7, independently from the adopted feature representation, the accuracy values are sensitive to varying λ. Fig.7 shows that choosing a value of $\lambda = 0.1$ always lead to similar or superior performance with respect to adopting a single-task clustering approach ($\lambda = 0$). The value $\lambda = 0.1$ correspond to the results reported in Fig.6. This clearly confirms the advantage of using a MTC approach for FPV analysis.

5 Conclusions

In this paper, we consider the problem of egocentric activity recognition from unlabeled data within a multi-task clustering framework. Two novel MTC algorithms have been proposed and evaluated extensively on two FPV datasets. Our experimental results clearly demonstrate the advantage of sharing informations among tasks over single tasks algorithms. Among our methods the approach based on EMD regularization achieves the best performance when used in its kernel version. On the other hand, our second algorithm is also effective as it is based on a convex optimization problem and it is particularly useful when one needs to incorporate some *a-priori* knowledge. In this paper we consider embedding information about temporal consistency but the CMTC method also permits to integrate *a-priori* knowledge about task dependencies if available (*e.g.* people performing the same activities in the same rooms correspond to more related tasks with respect to people operating in different rooms). This can be easily done by defining an appropriate matrix $\mathbf{B}$. Future works include exploiting the suitability of the proposed algorithms for other vision applications, as well as investigating how to improve our MTC methods (*e.g.* by detecting outlier tasks).

Acknowledgement. This work has been supported by the project Cluster Active Ageing at Home.

References

1. Kanade, T., Hebert, M.: First-person vision. Proceedings of the IEEE **100** (2012) 2442–2453
2. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: CVPR. (2013)
3. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR. (2012)
4. Ogaki, K., Kitani, K.M., Sugano, Y., Sato, Y.: Coupling eye-motion and ego-motion features for first-person activity recognition. In: CVPR Workshop on Egocentric Vision. (2012)
5. Tapia, E.M., Intille, S.S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors. In: Pervasive Computing. (2004) 158–175
6. Casale, P., Pujol, O., Radeva, P.: Human activity recognition from accelerometer data using a wearable device. In: Pattern Recognition and Image Analysis. Springer (2011) 289–296
7. Omid, A., Josephine, S., Stefan, C.: Novelty detection from an egocentric perspective. In: CVPR. (2011)
8. Fathi, A., Rehg, J.M.: Social interactions: A first-person perspective. In: CVPR. (2012)
9. Lei, J., Ren, X., Fox, D.: Fine-grained kitchen activity recognition using rgb-d. In: UBICOMP. (2012)
10. Taralova, E., De la Torre, F., Hebert, M.: Source constrained clustering. In: ICCV. (2011)
11. Mahasseni, B., Todorovic, S.: Latent multitask learning for view-invariant action recognition. In: ICCV. (2013)
12. Yan, Y., Liu, G., Ricci, E., Sebe., N.: Multi-task linear discriminant analysis for multi-view action recognition. In: ICIP. (2013)
13. Yuan, C., Hu, W., Tian, G., Yang, S., Wang, H.: Multi-task sparse learning with beta process prior for action recognition. In: CVPR. (2013)
14. Caruana, R.: Multitask learning. Machine learning **28** (1997) 41–75
15. Salakhutdinov, R., Torralba, A., Tenenbaum, J.: Learning to share visual appearance for multiclass object detection. In: CVPR. (2011)
16. Lu, G., Yan, Y., Sebe, N., Kambhamettu, C.: Knowing where i am: Exploiting multi-task learning for multi-view indoor image-based localization. In: BMVC. (2014)
17. Wang, X., Zhang, C., Zhang, Z.: Boosted multi-task learning for face verification with applications to web image and video search. In: CVPR. (2009)
18. Yan, Y., Ricci, E., Subramanian, R., Lanz, O., Sebe, N.: No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In: ICCV. (2013)
19. Gu, Q., Zhou, J.: Learning the shared subspace for multi-task clustering and transductive transfer classification. In: ICDM. (2009)
20. Kulis, B., Jordan, M.I.: Revisiting k-means: New algorithms via bayesian nonparametrics. In: ICML. (2012)
21. Zhang, J., Zhang, C.: Multitask bregman clustering. Neurocomput. **74** (2011) 1720–1734
22. Bulling, A., Ward, J.A., Gellersen, H., Troster, G.: Eye movement analysis for activity recognition using electrooculography. TPAMI **33** (2011) 741–753

23. Rubner, Y., Tomasi, C., Guibas, L.J.: A metric for distributions with applications to image databases. In: ICCV. (1998)
24. Borgwardt, K., Gretton, A., Rasch, M., Kriegel, H.P., Schoelkopf, B., Smola, A.: Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics **22** (2006) 1–9
25. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3** (2011) 1–122
26. Ding, C., Li, T., Jordan, M.I.: Convex and semi-nonnegative matrix factorizations. TPAMI **32** (2010) 45–55