Weakly Supervised Actor-Action Segmentation via Robust Multi-Task Ranking

Yan Yan[1], Chenliang Xu[2], Dawen Cai[3], Jason J. Corso[1]
[1]Department of Electrical Engineering and Computer Science, University of Michigan
[2]Department of Computer Science, University of Rochester
[3]Department of Cell and Developmental Biology, Biophysics, University of Michigan
{tomyan, dwcai, jjcorso}@umich.edu, {chenliang.xu}@rochester.edu

Abstract

Fine-grained activity understanding in videos has attracted considerable recent attention with a shift from action classification to detailed actor and action understanding that provides compelling results for perceptual needs of cutting-edge autonomous systems. However, current methods for detailed understanding of actor and action have significant limitations: they require large amounts of finely labeled data, and they fail to capture any internal relationship among actors and actions. To address these issues, in this paper, we propose a novel, robust multi-task ranking model for weakly-supervised actor-action segmentation where only video-level tags are given for training samples. Our model is able to share useful information among different actors and actions while learning a ranking matrix to select representative supervoxels for actors and actions respectively. Final segmentation results are generated by a conditional random field that considers various ranking scores for video parts. Extensive experimental results on the Actor-Action Dataset (A2D) demonstrate that the proposed approach outperforms the state-of-the-art weakly supervised methods and performs as well as the top-performing fully supervised method.

1. Introduction

Understanding fine-grained activities in videos is gaining attention in the video analysis community. Over the past decade, we have witnessed the shift of interest in the number of activities, e.g. from no more than ten [42, 29] to many hundreds [24, 5] and thousands [1]; in the scope of activities, e.g. from single person actions [45] to person-person interactions [43], person-object interactions [17], and even animal activities [19, 60]; and moreover, in the approaches to model activities, e.g. from classification [55, 53, 47] to localization [66, 49, 38, 46, 21], detection [12, 40, 8, 52] and segmentation [30, 36, 16]. The fine-grained results have also demonstrated their utilities in various emerging applications such as robot manipulation [41, 65] and video-and-language [48, 61].

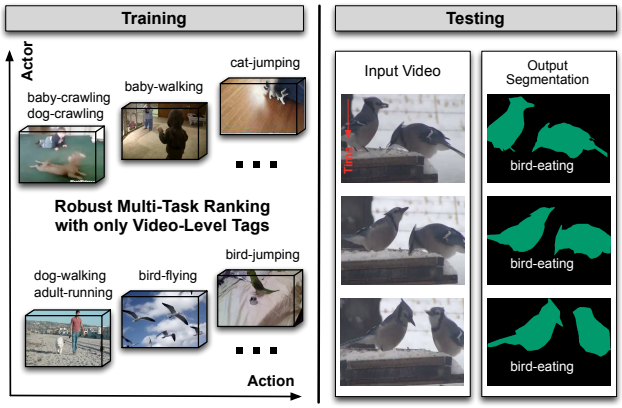

Figure 1. The weakly supervised actor-action semantic semgentation problem. Our method learns from weak supervision where only video-level tags for training videos are available, and generates pixel-level actor-action segmentation for a given testing video.

Among the many fine-grained activities, there is a growing interest in simultaneously understanding *actions* and *actors*, the agents who perform actions. It opens a new window to explore inter-agent and intra-agent activities for a comprehensive understanding. To address this issue, Xu et al. [60] introduced a new actor-action segmentation challenge on a difficult actor-action dataset (A2D), where they focused on spatiotemporal segmentation of seven types of actors, e.g. *human adult*, *dog* and *cat*, performing eight different actions, e.g. *walking*, *crawling*, *running*. In particular, the method proposed by Xu and Corso [58] sets the state of the art in this problem where they combine a labeling CRF with a supervoxel hierarchy to consider adaptive and long-ranging interactions among various actors performing various actions. Despite the success in pushing up the numbers in performance, their method together with many leading methods in activity segmentation [30, 36, 16] suffer largely from the following two aspects.

First, except Mosabbeb et al. [39], most methods in spa-

tiotemporal activity segmentation [60, 36, 58, 16, 30] are in a fully supervised setting where they require dense pixel-level annotation or bounding box annotation on many training samples. These assumptions are not realistic when we deal with real-world videos where available annotations are at most video-level tags or descriptions and have extreme diversity in the types of actors performing actions. Even humans alone can perform many hundreds of actions [6], not to mention the large variety in actors. Indeed, there are a few methods working on the problem of action co-segmentation [57, 16]. However, the ability to use weak supervision with only video-level tags for spatiotemporal activity segmentation is yet to be explored.

Second, existing methods in actor-action segmentation [60, 58] train classifiers independently for actors and actions, and only model their relationship in the random fields for segmentation output. Despite the success in considering different actor-action classification responses from various video parts, they lack the consideration of the interplay of actors and actions in features and classifiers, which is important as seen from the recent progress in image segmentation [35, 31]. For example, when separating the two fine-grained classes *dog-running* and *cat-running*, we should also benefit from extra information from all actions performed by the two actors.

To overcome the above limitations, we present a new robust multi-task ranking model that shares useful information among different actors and actions while learning a ranking matrix. The learned ranking matrix can be used for better potential generations due to this feature sharing. The regularization terms consist of a trace-norm and a $\ell_{1,2}$-norm, such that the model is able to capture a common set of features among relevant tasks and identify outlier tasks; hence, it is robust. We propose an efficient iterative optimization scheme for the problem. With this new learning model, we devise a pipeline to solve the weakly supervised actor-action segmentation problem where only video-level tags are given for the training videos (see Fig. 1). In particular, we first segment videos into supervoxels and extract features on supervoxels, then use the proposed robust multi-task ranking model to select representative supervoxels for actor and action respectively, and then use a CRF to generate the final segmentation output.

We conduct extensive experiments on the recently introduced large-scale A2D dataset [60]. In particular, we compare our methods against a set of fully supervised methods including the top-performing grouping process models [58]. For a comprehensive comparison, we also compare to a recent top-performing weakly supervised semantic segmentation method [54], and three learning methods including ranking SVM [23], dirty model multi-task learning [22], and clustered multi-task learning [70]. The experimental results show that our method outperforms all other weakly supervised methods and achieves performance as high as the top-performing fully supervised method.

2. Related Work

We have discussed the relationship of our method to existing actor-action segmentation methods in the introduction (Sec. 1). Recently, there are many emerging works on action detection [12, 40, 8, 52] and localization [66, 38, 49, 46, 21, 4]. We differ from them by considering pixel-level segmentation accuracy. Indeed, there are a few methods on spatiotemporal action segmentation [30, 36, 16, 39]. However, they all assume single type of actor and differ from our goal of actor-action segmentation.

Our work is also related to the many works in semantic video segmentation. Liu et al. [32] propose an object-augmented dense CRF in the spatio-temporal domain, which captures long-range dependencies between supervoxels and imposes consistency between object and supervoxel labels for multiclass video semantic segmentation. Kundu et al. [27] extend the fully connected CRF [26] to work on videos. Ladický et al. [28] build a hierarchical CRF on multi-scale segmentations that leverages higher-order potentials in inference. Despite the lack of explicit consideration of actors and actions, we compare to a representative subset of these methods [26, 28] in Sec. 5.

There are many weakly supervised video segmentation methods [68, 34, 51, 18] and co-segmentation methods [54, 11, 56, 67, 9]. Zhang et al. [68] propose a segmentation-by-detection framework to segment objects with video-level tags. Chiu et al. [9] study multi-class video co-segmentation where the number of object classes and number of instances at the frame and video level are unknown. Tsai et al. [54] propose an approach to segment objects and understand the visual semantics from a collection of videos that link to each other. However, these co-segmentation approaches lack any consideration of the internal relationship among different object categories, which is an important cue in the weakly-supervised segmentation approaches. In contrast, our framework is able to share useful information among different objects leading to better performance than the top-performing co-segmentation method [54] (see Sec. 5).

Multi-task learning has been effective in many applications, such as object detection [44] and classification [37, 62, 63, 64]. The idea is to learn models jointly that outperforms learning them separately for each task. To capture the task dependencies, a common approach is to constrain all the learned models to share a common set of features. This constraint motivates the introduction of a group sparsity term, i.e. the ℓ_1/ℓ_2-norm regularizer as in [2]. However, in practice, the ℓ_1/ℓ_2-norm regularizer may not be effective since not every task is related to all the others. To this end, the MTL algorithm based on the dirty model is proposed in [22] with the goal of identifying irrelevant (out-

lier) tasks. In some cases, the tasks exhibit a sophisticated group structure and it is desirable that the models of tasks in the same group are more similar to each other than to those from a different group. To model complex task dependencies, several clustered multi-task learning methods have been introduced [20, 69, 70]. Different from previous multi-task classification and regression problems, we propose a robust multi-task ranking model with the ability to identify outlier tasks. Meanwhile, an efficient solver is devised in this paper.

3. Robust Multi-Task Ranking

Our core technical emphasis builds on the current methods in learning a preference function for ranking, which has been widely used across fields [33]. To obtain good potentials for segmentation and select representative supervoxels and action tubes for specific categories (details in Sec. 4), we propose a robust multi-task ranking approach to share features among different actors and actions. In the rest of this section, we first give some background about SVM ranking, and then introduce our robust multi-task ranking.

Denote $\mathbf{x} \in I\!R^d$ as a d-dimensional feature vector and $\mathbf{w} \in I\!R^d$ as the learned weight parameter, the ranking SVM optimization problem is formulated as follows:

$$\begin{aligned} \min_{\mathbf{w},\varepsilon} \ & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum \varepsilon_{ij} \\ s.t. \ & \mathbf{w}^T\mathbf{x}_i \geq \mathbf{w}^T\mathbf{x}_j + 1 - \varepsilon_{ij} \\ & \varepsilon_{ij} \geq 0 \end{aligned} \tag{1}$$

where ε_{ij} are slack variables measuring the error of distance of the ranking pairs $(\mathbf{x}_i, \mathbf{x}_j)$. $\|\cdot\|$ is the ℓ_2-norm of a vector. The notation $(\cdot)^T$ indicates the transpose operator. C is the regularization parameter.

Given a set of related tasks, multi-task learning seeks to simultaneously learn a set of task-specific classification or regression models. The intuition behind multi-task learning is that a joint learning procedure accounting for task relationships is more efficient than learning each task separately. We first extend the ranking SVM to the multiple-task setting via the following optimization problem:

$$\begin{aligned} \min_{\mathbf{W},\gamma,\varepsilon} \ & \frac{1}{2}\|\mathbf{W}\|_F^2 + C_1\sum_{i,j\in S}\gamma_{ijk} + C_2\sum_{i,j\in D}\varepsilon_{ijk} + \lambda\Phi(\mathbf{W}) \\ s.t. \ & \left|\mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk}\right| \leq \gamma_{ijk} \\ & \mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk} \geq 1 - \varepsilon_{ijk} \\ & \varepsilon_{ijk} \geq 0 \\ & \gamma_{ijk} \geq 0 \end{aligned} \tag{2}$$

where $\mathbf{W} \in I\!R^{d\times K}$ is the learned ranking matrix as $[\mathbf{w}_1^T, \ldots, \mathbf{w}_k^T, \ldots, \mathbf{w}_K^T]$. $\mathbf{w}_k$ is the k-th column of $\mathbf{W}$. K is the number of tasks. C_1, C_2 and λ are regularization parameters. ε_{ijk} and γ_{ijk} are slack variables in the k-th task measuring the error of the distance between dissimilar pairs (i,j) in D satisfying $\mathbf{w}_i\mathbf{x}_i > \mathbf{w}_j\mathbf{x}_j$ and similar pairs (i,j) in S satisfying $\mathbf{w}_i\mathbf{x}_i \approx \mathbf{w}_j\mathbf{x}_j$. $\Phi(\mathbf{W})$ is the regularization term of $\mathbf{W}$.

The regularization term used in most traditional multi-task learning approaches assumes that all tasks are related [2] and their dependencies [20, 69, 70] can be modelled by a set of latent variables. However, in many real world applications, such as our actor-action semantic segmentation problem, not all tasks are related. When outlier tasks exist, enforcing erroneous and non-existent dependencies may lead to negative knowledge transfer. Take *actions* as an example, action tasks *climb, crawl, jump, roll, run, walk* may share useful information among each other, while the action task *eat* seems to be an outlier task. Incorporating *eat* in the multi-task learning may bring negative knowledge sharing.

In contrast, Chen et al. [7] propose regularization terms with a trace-norm plus a $\ell_{1,2}$-norm that simultaneously captures a common set of features among relevant tasks and identifies outlier tasks. They also theoretically proved a bound to measure how well the regularization terms approximate the underlying true evaluation. Inspired by them, we decompose our regularization term into two terms. One term enforces a trace norm on $\mathbf{L} \in I\!R^{d\times K}$ to encourage the desirable low-rank structure in the matrix to capture the shared features among different actions and actors. The other term enforces the group Lasso penalties on $\mathbf{E} \in I\!R^{d\times K}$ which induces the desirable group-sparse structure in the matrix to detect the outlier tasks. This formulation is robust to outlier tasks and effectively achieves joint feature learning based on the assumption that the same set of essential features are shared across different actions and actors with the existence of outlier tasks.

We hence propose the following optimization problem:

$$\begin{aligned} \min_{\mathbf{W},\gamma,\varepsilon} \ & \frac{1}{2}\|\mathbf{W}\|_F^2 + C_1\sum_{i,j\in S}\gamma_{ijk} + C_2\sum_{i,j\in D}\varepsilon_{ijk} \\ & + \lambda_1\|\mathbf{L}\|_* + \lambda_2\|\mathbf{E}\|_{1,2} \\ s.t. \ & \left|\mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk}\right| \leq \gamma_{ijk} \\ & \mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk} \geq 1 - \varepsilon_{ijk} \\ & \varepsilon_{ijk} \geq 0 \\ & \gamma_{ijk} \geq 0 \\ & \mathbf{W} = \mathbf{L} + \mathbf{E} \end{aligned} \tag{3}$$

In Eq. 3, the learned weighted matrix $\mathbf{W}$ is decomposed into $\mathbf{L} + \mathbf{E}$. The notation $\|\mathbf{L}\|_*$ = trace$(\sqrt{\mathbf{L}^*\mathbf{L}})$ is trace norm and $\|\mathbf{E}\|_{1,2} = \left[\sum_{j=1}^{K}(\sum_{i=1}^{d}|e_{ij}|)^2\right]^{1/2}$ is $\ell_{1,2}$-norm.

Although we adopt the same regularization term as [7], our proposed optimization is different in three critical aspects: (i) The optimization problem in [7] is a regression

problem while ours is a ranking optimization problem. This makes [7] unsuitable to be used in our actor-action video semantic segmentation with weakly supervised setting where good potentials for segmentation and representative supervoxels are needed. (ii) The loss function in [7] is a least-squared loss, which sometimes does not work well for real-world datasets because the least-squared loss has the tendency to be dominated by outliers. In our actor-action analysis, outlier tasks exist which further exaggerates this effect; (iii) The optimization method itself is different between [7] and our problem, as we explain next.

3.1. Optimization

The proposed optimization problem (Eq. 3) is hard to solve due to the mixture of different norms and constraints. To facilitate solving the original problem, we introduce a slack variable $\mathbf{S}$ to solve the optimization problem in an alternating way. The optimization problem can be decomposed into two separate steps by iteratively updating $\mathbf{W}$ and $\mathbf{S}$ respectively. With the slack variable, the optimization problem becomes:

$$\begin{aligned}\min_{\mathbf{W},\mathbf{S},\gamma,\varepsilon} \; & \frac{1}{2}\|\mathbf{W}\|_F^2 + C_1 \sum_{i,j\in S} \gamma_{ijk} + C_2 \sum_{i,j\in D} \varepsilon_{ijk} \\ & + \|\mathbf{W}-\mathbf{S}\|_F^2 + \lambda\Phi(\mathbf{S}) \\ s.t. \; & \left|\mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk}\right| \le \gamma_{ijk} \\ & \mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk} \ge 1-\varepsilon_{ijk} \\ & \varepsilon_{ijk} \ge 0 \\ & \gamma_{ijk} \ge 0 \end{aligned} \tag{4}$$

The term $\|\mathbf{W}-\mathbf{S}\|_F^2$ in Eq. 4 enforces the solution of $\mathbf{S}$ to be close to $\mathbf{W}$. The term $\Phi(\mathbf{S})$ is the regularization on $\mathbf{S}$. There are two major steps to optimize Eq. 4 as follows:
Step 1: Fix $\mathbf{S}$, optimize $\mathbf{W}$. Eq. 3 becomes,

$$\begin{aligned}\min_{\mathbf{w}_k,\gamma,\varepsilon} \; & \frac{1}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|^2 + C_1 \sum_{i,j\in S} \gamma_{ijk} + C_2 \sum_{i,j\in D} \varepsilon_{ijk} \\ & + \sum_{k=1}^{K}\|\mathbf{w}_k-\mathbf{s}_k\|^2 \\ s.t. \; & \left|\mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk}\right| \le \gamma_{ijk} \\ & \mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk} \ge 1-\varepsilon_{ijk} \\ & \varepsilon_{ijk} \ge 0 \\ & \gamma_{ijk} \ge 0 \end{aligned} \tag{5}$$

Eq. 5 can be decomposed into K separate single-task SVM ranking sub-problems and therefore can be solved via a standard SVM ranking solver [23].
Step 2: Fix $\mathbf{W}$, optimize $\mathbf{S}$. Eq. 3 becomes,

$$\min_{\mathbf{S}} \|\mathbf{S}-\mathbf{W}\|_F^2 + \lambda\Phi(\mathbf{S}) \tag{6}$$

Algorithm 1 Solving Eq. 4

INPUT: $\mathcal{D}_k$, $\mathcal{S}_k$, $\forall k = 1,\ldots,K$, λ_1, λ_2, C_1, C_2.
Initialize $\mathbf{W}_0$, $\mathbf{S}_0$.
LOOP:
1. Fix $\mathbf{S}$, optimize $\mathbf{W}$
   for k = 1 to K
      Fix $\mathbf{s}_k$, optimize Eq. 5 using [23], update $\mathbf{w}_k$
   end
2. Fix $\mathbf{W}$, optimize $\mathbf{S}$
   Optimize Eq. 6 using FISTA [3], update $\mathbf{S}$
Until Convergence
Output: W

The first term in Eq. 6 penalizes the learned slack weight matrix $\mathbf{S}$ to be close to the original matrix $\mathbf{W}$. This problem becomes a traditional multi-task learning problem and can be solved via the proximal gradient method FISTA [3]. The algorithm solving the proposed problem is summarized as in Algorithm 1.

4. Weakly Supervised Actor-Action Segmentation

In this section, we describe how we tackle the weakly supervised actor-action segmentation problem with our robust multi-task ranking model. The goal is to assign an actor-action label (e.g. *adult-eating* and *dog-crawling*) or a background label to each pixel in a video. We only have access to the video-level actor-action tags for the training videos. This problem is challenging as more than one-third of videos in A2D have multiple actors performing actions.

4.1. Overview

Figure 2 shows an overview of our framework. We first segment videos into supervoxels using the graph-based hierarchical supervoxel method (GBH) [14]. Meanwhile, we generate action tubes as the minimum bounding rectangles around supervoxels. We extract features at different GBH hierarchy levels to describe supervoxels and action tubes (see Sec. 4.2). Three different kinds of potentials (action, actor, actor-action) are computed via our robust multi-task ranking model by considering information sharing among different groups of actors and actions (see Sec. 4.3). Finally, we devise a CRF model for actor-action segmentation (see Sec. 4.4).

4.2. Supervoxels and Action Tubes

Supervoxels. Supervoxel segmentation defines a compact video representation where pixels in space-time with similar color and motion properties are grouped together. Various supervoxel methods are evaluated in [59]. Based on their work, we adopt the GBH supervoxel segmentation and con-

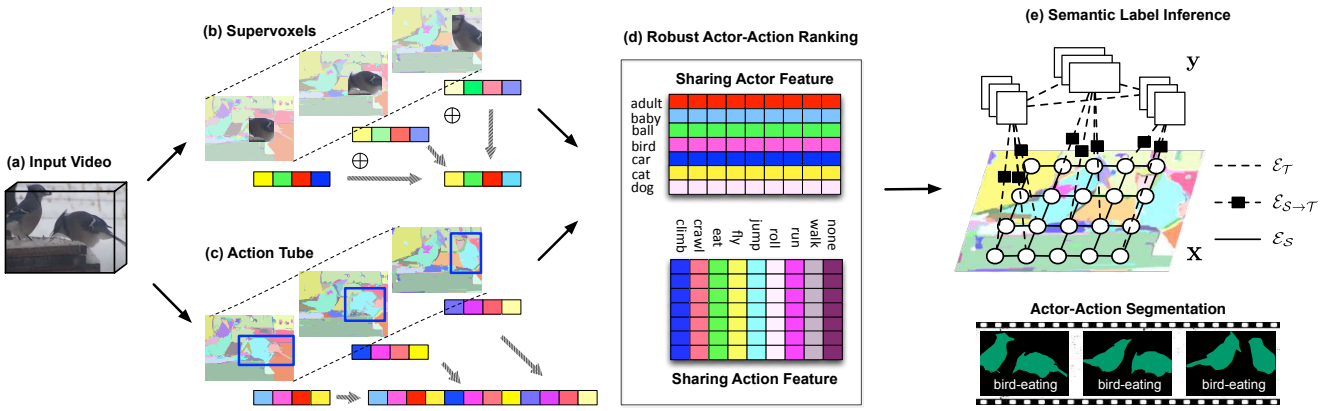

Figure 2. Overview of our proposed weakly supervised actor-action segmentation framework. (a) Input videos from the A2D dataset. (b) Supervoxel generation and feature extraction. (c) Action tube generation and feature extraction. (d) Sharing features among different actors and actions. (e) Semantic label inference for actor-action segmentation. Figure is best viewed in color and under zoom.

sider supervoxels from three different levels in a hierarchy. The performance of different levels are evaluated in Sec. 5. We extract CNN features from three time slices of a supervoxel, i.e. three superpixels, sampled from the beginning, the middle and the ending of supervoxel. We zero out pixels outside the superpixel boundary and use the rectangle image patch surrounding the superpixel as input to a pre-trained CNN to get *fc* vectors, similar to R-CNN [13]. The final feature vector representing the *actor* of a superpvoxel is averaged over the three time-slices as shown in Fig. 2 (b).

Action Tubes. Each supervoxel defines an action tube that is the sequence of minimum bounding rectangles around the supervoxel over time. Jain et al. [21] use such action tubes to localize human actions in videos. Here, we use them as proposals for general actions, e.g. *walking* and *crawling*, as well as fine-grained actor-actions, e.g. *cat-walking*, *dog-crawling*. We extract CNN features (*fc* vectors) from three sampled time slices of an action tube. The final feature vector representing *action* or *actor-action* of the action tube is a concatenation of the FC vectors as shown in Fig. 2 (c).

4.3. Robust Actor-Action Ranking

It is our assumption that information contained in supervoxel segments in *adult-running* videos should be correlated with supervoxel segments in *adult-walking* videos as they share same actor *adult*. Similarily, the correlation of action tubes among fine-grained actions in a same general action, e.g. *cat-walking* and *dog-walking*, should be larger than the correlation among non-relevant action pairs.

In the weakly supervised setting, we only have access to video-level tags for training videos. To better use this extremely weak supervision, we propose a robust multi-task ranking approach as described in Sec. 3 to effectively search for representative supervoxel segments and action tubes for each category and meanwhile, consider the sharing of useful information among different actors and actions. Three different sets of potentials (actor, action, actor-action) are obtained by sharing common features among tasks via the multi-task ranking approach by setting each task as action category (e.g. *walking*, *running* and *climbing*), actor category (e.g. *adult*, *cat* and *bird*) and actor-action category (e.g. *adult-walking*, *bird-climbing* and *car-rolling*).

4.4. Semantic Label Inference

We construct a CRF on the entire video. We denote $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ as a video with n supervoxels and define a set of random variables $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ on supervoxels, where x_i takes a label from the *actors*. Similarly, we denote $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ as a set of m action tubes and define a set of random variables $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ on action tubes, where y_i takes a label from the *actions*. A graph is constructed with three sets of edges: a set of edges $\mathcal{E}_\mathcal{S}$ linking neighboring supervoxels, a set of edges $\mathcal{E}_\mathcal{T}$ linking neighboring action tubes, and a set of edges $\mathcal{E}_{\mathcal{S}\to\mathcal{T}}$ linking supervoxels and action tubes. Our goal is to minimizes the following objective function:

$$(\mathbf{x}^*, \mathbf{y}^*) = \arg\min_{x,y} \sum_{(i,j)\in\mathcal{E}_\mathcal{S}} \psi(x_i, x_j) + \sum_{(i,j)\in\mathcal{E}_\mathcal{T}} \psi(y_i, y_j) + \sum_{i\in\mathcal{S}} \phi(x_i) + \sum_{i\in\mathcal{T}} \varphi(y_i) + \sum_{(i,j)\in\mathcal{E}_{\mathcal{S}\to\mathcal{T}}} \xi(x_i, y_j) \ , \quad (7)$$

where $\phi(\cdot)$, $\varphi(\cdot)$ and $\xi(\cdot)$ are the negative log of the normalized ranking scores for actor, action and actor-action respectively, and $\psi(\cdot,\cdot)$ takes the form of a contrast-sensitive Potts model to encourage smoothness. Following [58], we also use video-level potentials as an additional global labeling cost. Comparing to the models in [60], our model is more flexible and allows separate topologies for supervoxels and action tubes (see Fig. 2 (e)). Finally the segmentation is generated by mapping action tubes to supervoxels.

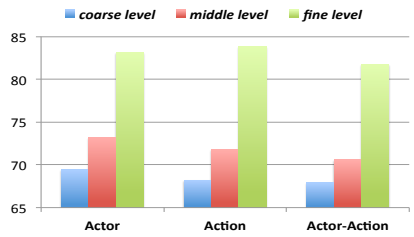

Figure 3. The overall pixel accuracy for different GBH hierarchy supervoxels. Figure is best viewed in color.

5. Experiments

We perform extensive experiments on the A2D dataset to evaluate our proposed method for weakly supervised actor-action segmentation. We first describe our experimental settings, and then present our results.

Dataset. Fine-grained actor-action segmentation is a newly proposed problem. To the best of our knowledge, there is only one actor-action video dataset, i.e. A2D [60], in literature. The A2D dataset contains 3782 videos that are collected from YouTube. Both the pixel-level labeled actors and actions are available with the released dataset. The dataset includes eight different actions, e.g. *climbing, crawling, eating, flying, jumping, rolling, running, walking*, and one additional *none* action. The *none* action class means that the actor is not performing an action or is performing an action that is outside their consideration. Meanwhile, seven actor classes, e.g. *adult, baby, ball, bird, car, cat, dog*, are considered in A2D to perform those actions.

Experimental Settings. We use GBH [15] to generate hierarchical supervoxel segmentations. We evaluate our method on three GBH hierarchy levels (fine, middle, coarse) where the number of supervoxels varies from 20-200 in each video. The action tubes are generated with minimum bounding rectangles around supervoxels. For supervoxel and action tube features, we use pretained GoogLeNet [50] to extract CNN deep features of the average pooling layer 1024-dimensional feature vector. GoogLeNet is a 22-layer deep network which has achieved good performance in the context of image classification and object detection. The regularization parameters λ_1, λ_2 and C_1, C_2 are grid-searched via range $[0.01, 0.1, 1, 10, 100]$ for training our robust multi-task ranking model. We use multi-label graph cuts [10] for CRF inference and empirically set the parameters by hand. We follow the same setup as [60] for the training/testing split of the dataset.

Evaluation Metrics. For actor-action segmentation, pixel-level accuracy is the most commonly used measurement in literature. We use two metrics in the paper: (i) The Overall Pixel accuracy measures the proportion of correctly labeled pixels to all pixels in ground-truth frames. (ii) The Per-Class accuracy measures the proportion of correctly labeled pixels for each class and then averages over all classes.

Table 1. Comparison of overall pixel accuracy on the A2D dataset.

	Action	Actor	Actor-Action
AHRF [28]	63.9	64.9	63.0
GPM [58]	82.4	82.2	80.8
FCRF [25]	77.6	77.9	76.2
RSVM [23]	70.1	70.8	68.8
DM-MTL [22]	72.3	72.9	71.4
C-MTL [70]	73.1	73.5	72.7
WSS [54]	71.5	71.9	70.4
Ours	**83.8**	**83.1**	**81.7**

5.1. Comparison to Variations of Our Method

We evaluate our approach with different GBH hierarchy supervoxels. The overall pixel accuracy of segmentation results are shown in Fig. 3. We observe that the fine-level GBH hierarchy achieves considerably better results than coarser-level GBH hierarchies. This is probably because fine-level GBH hierarchy has a reasonable number of supervoxels (100-200) for each video, which leads to the best raw segmentation result among the three. We use fine-level GBH hierarchy supervoxels in the rest of our experiments.

We also perform experiments to show the impact of different types of potentials used. We achieve 81.7% overall pixel accuracy when we use both coarse labels (actor and action) and fine-grained labels (actor-action), and 72.6% overall pixel accuracy when we use only fine-grained labels. In the latter case, a simple pairwise CRF is constructed for action tubes. The results support the explicit consideration of information sharing among fine-grained actions.

5.2. Comparison to State-of-The-Art Methods

We compare our method to state-of-the-art fully supervised segmentation methods, such as Associate Hierarchical Random Fields (AHRF) [28], Grouping Process Models (GPM) [58], and Fully-Connected CRF (FCRF) [25]. Since our method is in the weakly supervised setting, we also compare against a recently published top-performing method in weakly supervised semantic video segmentation (WSS) [54]. For a comprehensive understanding, we also compare our robust multi-task ranking model with other learning models, including single-task learning and multi-task learning approaches, such as Ranking SVM (RSVM), Dirty Model Multi-Task Learning (DM-MTL) [22], and Clustered Multi-Task Learning (C-MTL) [70]. For fair comparison, we use author-released code for methods [58, 54]. For Ranking SVM, we use the released implementation in [23]. For multi-task learning approaches [22, 70], we use the MALSAR toolbox [71]. We use the same experiment setup as ours for the learning models and weakly supervised method. Notice that the fully supervised methods have access to pixel-level annotation for the training videos.

Table 1 shows the overall pixel accuracy for all methods. We observe that our method outperforms all other base-

Table 2. Comparison of per-class accuracy on the A2D dataset (top-2 scores for each category are highlighted).

		baby					ball				car				
method	BK	climb	crawl	roll	walk	none	fly	jump	roll	none	fly	jump	roll	run	none
AHRF [28]	69.2	21.3	5.5	39.8	13.5	0.0	3.2	2.3	13.6	1.5	18.1	68.0	13.6	47.9	**12.2**
GPM [58]	**88.4**	**65.4**	**65.0**	58.4	**61.5**	0.0	11.3	**28.3**	21.1	0.0	**41.2**	**86.3**	**70.9**	**65.9**	0.0
FCRF [25]	82.2	3.4	23.4	41.0	17.8	0.0	3.7	0.3	1.0	0.0	13.7	**78.4**	**55.4**	43.7	1.8
RSVM [23]	72.7	0.1	5.5	67.8	3.8	1.2	4.0	5.7	12.5	1.6	14.8	30.4	37.8	37.7	5.3
DM-MTL [22]	83.0	51.8	50.1	58.3	47.9	0.0	9.4	11.7	16.6	0.0	33.2	64.9	42.3	47.4	0.0
C-MTL [70]	**83.0**	49.0	61.9	**75.4**	40.9	**28.8**	**19.5**	16.3	33.4	**13.2**	30.9	36.4	32.5	38.8	7.0
WSS [54]	74.1	16.0	10.9	50.9	21.9	7.9	4.0	5.0	**49.2**	1.7	17.8	52.4	13.5	35.1	5.2
Ours	82.2	**66.2**	**73.6**	**78.5**	**52.5**	**33.5**	**19.5**	**20.1**	**62.6**	**13.2**	**46.2**	65.6	42.5	**49.4**	**22.7**

	adult								bird						
method	climb	crawl	eat	jump	roll	run	walk	none	climb	eat	fly	jump	roll	walk	none
AHRF [28]	0.0	56.0	6.1	1.1	0.0	0.0	15.3	10.9	14.6	11.4	19.9	5.0	29.6	7.5	0.0
GPM [58]	**74.8**	**81.0**	**76.4**	**49.3**	**52.4**	**50.4**	41.0	0.0	**60.6**	**38.8**	**66.5**	**17.5**	**45.9**	**47.9**	0.0
FCRF [25]	21.6	**64.5**	46.3	25.3	12.0	**50.9**	26.9	33.8	25.9	16.1	**57.3**	17.1	35.0	7.4	0.0
RSVM [23]	2.9	27.9	41.2	1.7	2.9	10.0	7.6	**57.2**	9.0	1.0	39.8	1.1	43.2	14.9	0.0
DM-MTL [22]	44.5	43.9	67.1	27.7	34.5	35.3	32.7	0.0	47.7	27.4	51.3	13.6	32.1	30.4	0.0
C-MTL [70]	38.5	38.4	69.4	28.8	46.6	27.4	**41.0**	46.5	26.5	**27.7**	55.4	**45.0**	**60.2**	**36.9**	**6.0**
WSS [54]	6.6	23.5	50.8	9.6	10.1	11.1	15.3	29.0	33.6	14.5	30.1	8.2	31.1	21.0	0.0
Ours	**44.9**	47.8	**74.7**	**33.9**	**49.2**	42.1	**46.3**	**53.1**	**47.7**	27.4	51.3	13.6	32.1	30.4	0.0

	dog							cat							**Avg**
method	crawl	eat	jump	roll	run	walk	none	climb	eat	jump	roll	run	walk	none	-
AHRF [28]	13.2	16.4	0.0	0.0	0.0	0.0	0.0	18.3	38.8	0.0	8.8	0.0	9.3	0.0	13.9
GPM [58]	44.1	61.5	**31.4**	**62.6**	25.7	**74.2**	0.0	**42.8**	**52.3**	**33.7**	**71.7**	**48.0**	19.1	0.0	**43.9**
FCRF [25]	11.7	35.7	2.2	31.9	25.2	40.2	0.0	25.3	33.6	2.5	33.9	**48.9**	21.5	0.8	25.4
RSVM [23]	3.7	33.6	5.7	24.2	0.6	9.7	0.0	5.0	38.6	0.2	43.8	0.0	5.6	0.1	16.7
DM-MTL [22]	36.9	65.6	26.9	50.9	22.2	59.8	0.0	16.9	46.5	12.1	66.2	25.6	7.7	0.0	32.8
C-MTL [70]	**45.5**	**80.9**	24.6	57.3	**37.7**	42.8	**3.6**	23.6	52.1	22.1	68.9	24.2	**39.1**	**23.1**	38.9
WSS [54]	16.2	36.3	10.3	24.3	1.0	18.4	1.4	13.6	42.0	8.2	46.3	0.5	15.8	0.3	20.3
Ours	**64.5**	**85.7**	**50.1**	**72.3**	**68.5**	**61.1**	**7.6**	**41.4**	**72.9**	**36.6**	**86.2**	36.7	**65.1**	**25.5**	**41.7**

lines. Our approach has 11% higher accuracy than the other weakly supervised approach (WSS) [54]. Their approach is unable to share feature similarity among different actions and actors which is very important in the weakly-supervised setting. Moreover, our method outperforms other single task learning (RSVM) and multi-task learning (DM-MTL, C-MTL) approaches by up to 20%, 9%, 3% respectively, which shows the robustness of our approach. Table 2 shows the per-class accuracy for all actor-action pairs on the A2D dataset. We observe that our approach outperforms all other baselines in averaged performance except GPM [58]. However, we note that GPM is a fully supervised approach, i.e. it needs tedious pixel-level human labelling for training samples. In addition, our method works well on the actor categories 'dog' and 'cat' which shows the ability of our method to identify outlier tasks to better share features among different tasks.

Figure 4 shows qualitative results of our approach and other methods. We observe that our approach can generate better visual qualitative results than other approaches. However, our method fails in some cases, such as *cat-jumping*. This is probably because there are several cats jumping simutaneously and motion is significant in the video.

6. Conclusion and Future Work

In this paper, we propose a novel weakly supervised actor-action segmentation method. In particular, a robust multi-task ranking model is devised to select the most representative supervoxels and action tubes for actor, action and actor-action respectively. Features are shared among different actors and actions via multi-task learning by simultaneously detecting outlier tasks. A CRF model is used for semantic label inference. The extensive experiments on the large-scale A2D dataset show the effectiveness of our proposed approach. One drawback of our apporach is that the ranking weights are learned independent from feature extraction in our framework. Future work includes exploring the possibility of using CNNs for actor-action analysis, such as multi-task learning with CNNs or FCN [35] for actor-action segmentation.

Acknowledgement. This work has been supported in part by a University of Michigan MiBrain grant, Google, Samsung, DARPA W32P4Q-15-C-0070 and ARO W911NF-15-1-0354.

Figure 4. Qualitative results shown in sampled frames for several video sequences from the A2D dataset. Columns from left to right are input video, ground-truth, our method, GPM [58], WSS [54], RSVM [23], DM-MTL [22] and AHRF [28] respectively. Our method is able to generate correct actor-action segmentation expect for *cat-jumping* and *adult-running* in these examples.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. Technical report, arXiv preprint arXiv:1609.08675, 2016. 1

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, 2007. 2, 3

[3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Science*, 2(1):183–220, 2009. 4

[4] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Weakly supervised action labeling in videos under ordering constraints. In *ECCV*, 2014. 2

[5] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 1

[6] Y.-W. Chao, Z. Wang, R. Mihalcea, and J. Deng. Mining semantic affordances of visual object categories. In *CVPR*, 2015. 2

[7] J. Chen, J. Zhou, and J. Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *ACM SIGKDD Conferences on Knowledge Discovery and Data Mining*, 2011. 3, 4

[8] W. Chen and J. J. Corso. Action detection by implicit intentional motion clustering. In *ICCV*, 2015. 1, 2

[9] W.-C. Chiu and M. Fritz. Multi-class video co-segmentation with a generative multi-video model. In *CVPR*, 2013. 2

[10] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. *International journal of computer vision*, 96(1):1–27, 2012. 6

[11] H. Fu, D. Xu, B. Zhang, and S. Lin. Object-based multiple foreground video co-segmentation. In *CVPR*, 2014. 2

[12] R. D. Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek, and T. Tuytelaars. Online action detection. In *ECCV*, 2016. 1, 2

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, 2016. 5

[14] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, 2010. 4

[15] M. Grundmann, V. Kwatra, M. Han, and E. I. Efficient hierarchical graph-based video segmentation. In *CVPR*, 2010. 6

[16] J. Guo, Z. Li, L.-F. Cheong, and S. Z. Zhou. Video co-segmentation for meaningful action extraction. In *ICCV*, 2013. 1, 2

[17] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009. 1

[18] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra, O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, and R. Sukthankar. Weakly supervised learning of object segmentations from web-scale video. In *ECCV Workshops*, pages 198–208. Springer, 2012. 2

[19] Y. Iwashita, A. Takamine, R. Kurazume, and M. S. Ryoo. First-person animal activity recognition from egocentric videos. In *IEEE International Conference on Pattern Recognition*, 2014. 1

[20] L. Jacob, F. Bach, and J. Vert. Clustered multi-task learning: A convex formulation. In *NIPS*, 2008. 3

[21] M. Jain, J. Van Gemert, H. Jégou, P. Bouthemy, C. Snoek, et al. Action localization with tubelets from motion. In *CVPR*, 2014. 1, 2, 5

[22] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *NIPS*, 2010. 2, 6, 7, 8

[23] T. Joachims. Training linear svms in linear time. In *ACM SIGKDD Conferences on Knowledge Discovery and Data Mining*, 2006. 2, 4, 6, 7, 8

[24] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1

[25] P. Krahenbuhl and V. Keltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 6, 7

[26] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 2

[27] A. Kundu, V. Vineet, and V. Koltun. Feature space optimization for semantic video segmentation. In *CVPR*, 2016. 2

[28] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative hierarchical random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1056–1077, 2014. 2, 6, 7, 8

[29] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1

[30] C. Lea, A. Reiter, R. Vidal, and G. D. Hager. Segmental spatiotemporal cnns for fine-grained action segmentation. In *ECCV*, 2016. 1, 2

[31] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016. 2

[32] B. Liu and X. He. Multiclass semantic video segmentation with object-level active inference. In *CVPR*, 2015. 2

[33] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009. 3

[34] X. Liu, D. Tao, M. Song, Y. Ruan, C. Chen, and J. Bu. Weakly supervised multiclass video segmentation. In *CVPR*, 2014. 2

[35] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2, 7

[36] J. Lu, R. Xu, and J. J. Corso. Human action segmentation with hierarchical supervoxel consistency. In *CVPR*, 2015. 1, 2

[37] Y. Luo, D. Tao, B. Geng, C. Xu, and S. Maybank. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE Transactions on Transactions on Pattern Recognition and Machine Intelligence*, 22(2):523–536, 2013. 2

[38] P. Mettes, J. C. van Gemert, and C. G. Snoek. Spot on: Action localization from pointly-supervised proposals. In *ECCV*, 2016. 1, 2

[39] E. A. Mosabbeb, R. Cabral, F. De la Torre, and M. Fathy. Multi-label discriminative weakly-supervised human activity recognition and localization. In *Asian Conference on Computer Vision*, 2014. 1, 2

[40] X. Peng and C. Schmid. Multi-region two-stream r-cnn for action detection. In *ECCV*, 2016. 1, 2

[41] L. Pinto, D. Gandhi, Y. Han, Y.-L. Park, and A. Gupta. The curious robot: Learning visual representations via physical interactions. In *ECCV*, 2016. 1

[42] M. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 1

[43] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009. 1

[44] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011. 2

[45] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *IEEE International Conference on Pattern Recognition*, 2004. 1

[46] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016. 1, 2

[47] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1

[48] Y. C. Song, I. Naim, A. Al Mamun, K. Kulkarni, P. Singla, J. Luo, D. Gildea, and H. Kautz. Unsupervised alignment of actions in video with text descriptions. In *International Joint Conference on Artificial Intelligence*, 2016. 1

[49] K. Soomro, H. Idrees, and M. Shah. Predicting the where and what of actors and actions through online action localization. In *CVPR*, 2016. 1, 2

[50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 6

[51] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei. Discriminative segment annotation in weakly labeled video. In *CVPR*, 2013. 2

[52] Y. Tian, R. Sukthankar, and M. Shah. Spatiotemporal deformable part models for action detection. In *CVPR*, 2013. 1, 2

[53] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 1

[54] Y.-H. Tsai, G. Zhong, and M.-H. Yang. Semantic co-segmentation in videos. In *ECCV*, 2016. 2, 6, 7, 8

[55] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 1

[56] L. Wang, G. Hua, R. Sukthankar, J. Xue, and N. Zheng. Video object discovery and co-segmentation with extremely weak supervision. In *ECCV*, 2014. 2

[57] C. Xiong and J. J. Corso. Coaction discovery: segmentation of common actions across multiple videos. In *ACM International Workshop on Multimedia Data Mining*, 2012. 2

[58] C. Xu and J. J. Corso. Actor-action semantic segmentation with grouping process models. In *CVPR*, 2016. 1, 2, 5, 6, 7, 8

[59] C. Xu and J. J. Corso. Libsvx: A supervoxel library and benchmark for early video processing. *International Journal of Computer Vision*, 119(3):272–290, 2016. 4

[60] C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso. Can humans fly? action understanding with multiple classes of actors. In *CVPR*, 2015. 1, 2, 5, 6

[61] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1

[62] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *ICCV*, 2013. 2

[63] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, and N. Sebe. A multi-task learning framework for head pose estimation under target motion. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 38(6):1070–1083, 2016. 2

[64] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe. Multi-task linear discriminant analysis for multi-view action recognition. *IEEE Transactions on Image Processing*, 23(12):5599–5611, 2014. 2

[65] Y. Yang, Y. Li, C. Fermüller, and Y. Aloimonos. Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In *AAAI Conference on Artificial Intelligence*, 2015. 1

[66] J. Yuan, B. Ni, X. Yang, and A. A. Kassim. Temporal action localization with pyramid of score distribution features. In *CVPR*, 2016. 1, 2

[67] D. Zhang, O. Javed, and M. Shah. Video object co-segmentation by regulated maximum weight cliques. In *ECCV*, 2014. 2

[68] Y. Zhang, X. Chen, J. Li, C. Wang, and C. Xia. Semantic object segmentation via detection in weakly labeled video. In *CVPR*, 2015. 2

[69] Y. Zhang and D. Yeung. A convex formulation for learning task relationships in multi-task learning. In *Uncertainty in Artificial Intelligence*, 2010. 3

[70] J. Zhou, J. Chen, and J. Ye. Clustered multi-task learning via alternating structure optimization. In *NIPS*, 2011. 2, 3, 6, 7

[71] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011. 6