International Journal of Computer Vision (2020) 128:1414–1432
https://doi.org/10.1007/s11263-019-01244-7

A Weakly Supervised Multi-task Ranking Framework for Actor–Action Semantic Segmentation

Yan Yan[1] · Chenliang Xu[2] · Dawen Cai[3] · Jason J. Corso[4]

Received: 27 September 2018 / Accepted: 24 September 2019 / Published online: 28 October 2019

Abstract
Modeling human behaviors and activity patterns has attracted significant research interest in recent years. In order to accurately model human behaviors, we need to perform fine-grained human activity understanding in videos. 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 Schatten p-norm 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 both the actor–action dataset and the Youtube-objects dataset 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.

Keywords Weakly supervised learning · Actor–action semantic segmentation · Multi-task ranking

1 Introduction

Observing people and trying to predict what they will perform next can provide a real learning experience. Human behavior is quite predictable in many instances. Behaviors can be extremely complex but there are areas that can be understood with a high degree of accuracy. Human behavior understanding has attracted significant research interest in recent years. To accurately model human behaviors, we need to perform fine-grained human activity understanding in videos. In this paper, we propose an approach that is able to generate fine-grained actor–action video semantic segmentation maps which can be further used for behavior understanding. After segmentation of actors in video sequences, the next step is to recognize and understand the behaviors of actors. The essence of behavior understanding may be considered to be a classification problem towards time varying data. Accordingly, two critical issues need to be addressed during classification. The first one is to obtain the reference behavior sequences and the other one needs to enable the training and matching methods effec-

Communicated by Xavier Alameda-Pineda, Elisa Ricci, Albert Ali Salah, Nicu Sebe, Shuicheng Yan.

✉ Yan Yan
y_y34@txstate.edu

✉ Dawen Cai
dwcai@umich.edu

Chenliang Xu
chenliang.xu@rochester.edu

Jason J. Corso
jjcorso@umich.edu

[1] Department of Computer Science, Texas State University, San Marcos, USA

[2] Department of Computer Science, University of Rochester, Rochester, USA

[3] Department of Cell and Developmental Biology, Biophysics, University of Michigan, Ann Arbor, USA

[4] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, USA

tive to cope with the minor deviation in both temporal and spatial scales for similar motion patterns. 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 (Rodriguez et al. 2008; Laptev et al. 2008) to many hundreds (Karpathy et al. 2014; Caba Heilbron et al. 2015) and thousands (Abu-El-Haija et al. 2016); in the scope of activities, e.g. from single person actions (Schuldt et al. 2004) to person–person interactions (Ryoo and Aggarwal 2009), person–object interactions (Gupta et al. 2009), and even animal activities (Iwashita et al. 2014; Xu et al. 2015); and moreover, in the approaches to model activities, e.g. from classification (Wang and Schmid 2013; Tran et al. 2015; Simonyan and Zisserman 2014) to localization (Jain et al. 2014; Yuan et al. 2016; Soomro et al. 2016; Mettes et al. 2016; Shou et al. 2016), detection (Geest et al. 2016; Peng and Schmid 2016; Chen and Corso 2015; Tian et al. 2013) and segmentation (Lea et al. 2016; Lu et al. 2015; Guo et al. 2013). The fine-grained results have also demonstrated their utilities in various emerging applications such as robot manipulation (Pinto et al. 2016; Yang et al. 2015) and video-and-language (Song et al. 2016; Xu et al. 2016).

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. (2015) 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 (2016a) sets the state-of-the-art in this problem where they combine a labeling Conditional Random Field 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 (Lea et al. 2016; Lu et al. 2015; Guo et al. 2013) suffer largely from the following two aspects.

First, except (Mosabbeb et al. 2014), most methods in spatiotemporal activity segmentation (Xu et al. 2015; Lu et al. 2015; Xu and Corso 2016a; Guo et al. 2013; Lea et al. 2016) 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 (Chao et al. 2015), not to mention the large variety in actors. Indeed, there are a few methods working on the problem of action co-segmentation (Xiong and Corso 2012; Guo et al. 2013). 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 (Xu et al. 2015; Xu and Corso 2016a) train classifiers independently for actors and actions, and only model their relationship in 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 (Long et al. 2015; Lin et al. 2016). 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. In many real-world applications involving multiple tasks, it is usually the case that a group of tasks are related while some other tasks are irrelevant to such a group. Simply pooling all tasks together and learning them simultaneously under a presumed structure may degrade the overall learning performance. Identifying irrelevant (outlier) tasks while learning multiple tasks referred as *robust multi-task learning* (Yu et al. 2007). In our previous work (Yan et al. 2017), we performed a trace-norm and a $\ell_{1,2}$-norm to capture a common set of features among relevant tasks and identify outlier tasks. Although the trace-norm minimization based objective is a convex problem with global solution, the relaxation may make the solution seriously deviate from the original solution. It is desired to solve a better approximation of the rank minimization problem without introducing much computational cost. This paper proposed a more flexible regularization Schatten p-norm term in the objective function. The regularization terms consist of a Schatten p-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 Conditional Random Field (CRF)

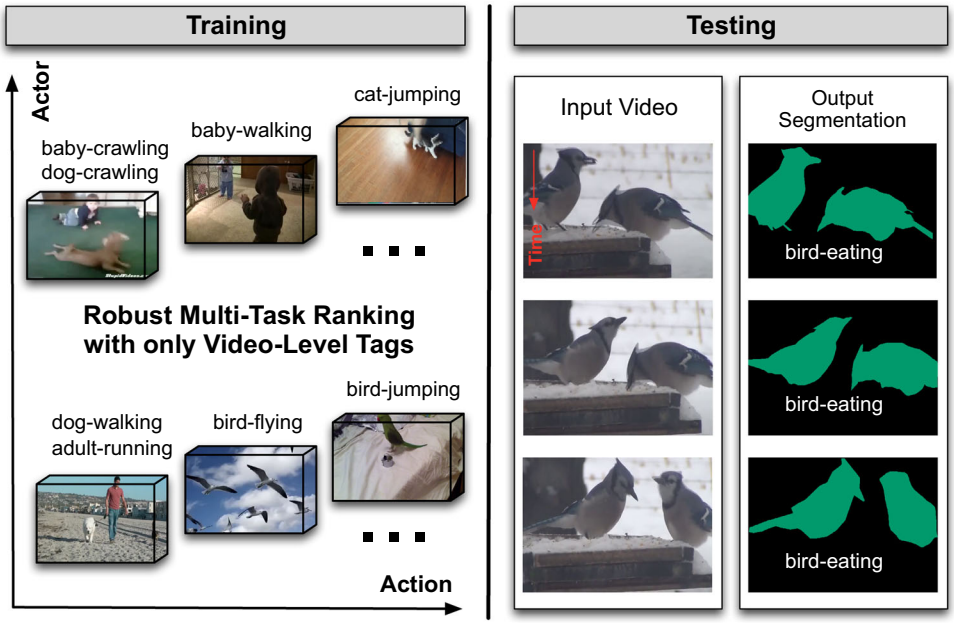

Fig. 1 The weakly supervised actor–action semantic semgentation problem. Our method learns from weak supervision where only video-level tags for training videos are available, and generates pixel-level actor–action segmentation for a given testing video

to generate the final segmentation output. Each supervoxel belongs to one or more parts of actors or scenes, which are quite different in terms of the contents (e.g. usually roads are smooth and actors are textured). To understand the contents of each supervoxel, we first collect all the supervoxels in videos with such label for each semantic category. We then select representative supervoxels through ranking SVM. These representative supervoxels selected in each category are further utilized in CRF, in which we assign each supervoxel a potential to be a specific category.

We conduct extensive experiments on the recently introduced large-scale A2D dataset (Xu et al. 2015) and Youtube-objects dataset (Prest et al. 2012). In particular, we compare our methods against a set of fully supervised methods including the top-performing grouping process models (Xu and Corso 2016a). For a comprehensive comparison, we also compare to a recent top-performing weakly supervised semantic segmentation method (Tsai et al. 2016), and other learning methods including ranking SVM (Joachims 2006), dirty model multi-task learning (Jalali et al. 2010), and clustered multi-task learning (Zhou et al. 2011a). 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.

To summarize, the main contributions of this paper are: (i) a pipeline is proposed to solve the weakly supervised actor–action segmentation problem where only video-level tags are given for the training videos; (ii) a new Schatten p-norm robust multitask ranking model, which shares useful information among different actors and actions while learning a ranking matrix, is presented; (iii) an efficient iterative optimization scheme for the Schatten p-norm robust multitask ranking problem is devised.

The paper is organized as follow. Section 2 reviews related work. Section 3 describes the Schatten p-norm robust multi-task ranking model. Section 4 introduces our approach for weakly supervised actor–action segmentation. Experiments are presented in Sect. 5, and conclusion is stated in Sect. 6.

2 Related Work

In this section, we review the related work from perspectives of video segmentation, semantic segmentation, co-localization, actor–action segmentation and multi-task learning and ranking, respectively.

2.1 Video Segmentation

Video segmentation is a fundamental and emerging topic in computer vision which potentially can be used for different applications, such as action and activity recognition, large-scale video retrieval, video event detection. In literature, video segmentation can leverage information from appearance (Brendel and Todorovic 2009; Grundmann et al. 2010), motion (Brox and Malik 2010) and multiple cues (Galasso et al. 2012). Different approaches have been used for video segmentation, such as generative layered approach (Kumar et al. 2005), graph-based approach (Grundmann et al. 2010), mean-shift approach (Paris 2008), manifold-embedding approaches (Brox and Malik 2010; Galasso et al. 2012). In particular, Xu and Corso (2012) evaluated different supervoxel methods for video segmentation, such as

segmentation by weight aggregation (SWA) (Corso et al. 2008), graph-based (GB) (Felzenszwalb and Huttenlocher 2004), hierarchical graph-based (GBH) (Grundmann et al. 2010). They identified GBH and SWA as the most effective supervoxel methods based on several generic and application independent criteria. There are many challenges for video segmentation. One major difficulty is the burden of labelling training samples, making the video segmentation unsolved. Due to this reason, most video segmentation approaches in literature are in unsupervised settings. However, unsupervised approaches usually perform not well and are computational expensive. To address these issue, different from previous unsupervised approaches, our approach leverage video-level label information which prevent us from tedious labelling work for video segmentation.

2.2 Semantic Segmentation

Semantic segmentation has attracted attention recently in computer vision. Some deep learning approaches have been proposed for image semantic segmentation, such as the famous Fully Convolutional Networks (FCN) (Long et al. 2015). Further, Zheng et al. (2015) introduced a form of convolutional neural network that combines the strengths of Convolutional Neural Networks (CNNs) and Conditional Random Fields (CRFs)-based probabilistic graphical modelling for image semantic segmentation. However, these approaches are not suitable for video semantic segmentation partially due to lack of training data and complexity of the video segmentation problem. For video semantic segmentation, few work has been done in literature. Some existing works addressed temporal coherence of pixel labelling (Lezama et al. 2011; Liu and He 2015). Lezama et al. (2011) used optical flow based long-term trajectories to discover moving objects. Liu and He (2015) proposed an object-augmented dense CRF in spatio-temporal domain, which captured long-range dependency between supervoxels, and imposed consistency between object and supervoxel labels for multiclass video semantic segmentation. For actor–action video semantic segmentation, Xu and Corso (2016a) proposed a grouping process model that combined local labelling CRFs with a hierarchical supervoxel decomposition. The supervoxels provided cues for possible groupings of nodes at various scales in the CRFs to encourage adaptive, high-order groups for more effective labelling.

2.3 Co-localization

Co-localization is a kind of weakly supervised localization approach (Deselaers et al. 2012) where strong supervision is not needed. Tang et al. (2014) proposed a co-localization approach via combining an image model and box model into a joint optimization problem. Joulin et al. (2014) introduced a formulation for video co-localization that is able to naturally incorporate temporal consistency in a quadratic programming framework. However, co-localization approaches overlooked the semantic meaning from superpixels/supervoxels which prevent them to be used for image and video semantic segmentation.

2.4 Actor–Action Segmentation

Recently, there are many emerging works on action detection (Geest et al. 2016; Peng and Schmid 2016; Chen and Corso 2015; Tian et al. 2013) and localization (Yuan et al. 2016; Mettes et al. 2016; Soomro et al. 2016; Shou et al. 2016; Jain et al. 2014; Bojanowski et al. 2014). We differ from them by considering pixel-level segmentation accuracy. There are only a few methods on spatiotemporal action segmentation (Lea et al. 2016; Lu et al. 2015; Guo et al. 2013; Mosabbeb et al. 2014). However, they all assumed single type of actor and differ from our goal of actor–action segmentation. The actor–action segmentation problem was first introduced in Xu et al. (2015), where a set of CRFs was proposed to consider various actor–action interactions in labeling supervoxels. Later, Xu and Corso (2016a) presented a grouping process model that combined local labelling CRFs with a supervoxel hierarchy. The supervoxel hierarchy provided cues for possible groupings of nodes at various scales in the CRFs to encourage adaptive long-ranging interactions. This method sets the state-of-the-art on the A2D dataset. In line with our work, there are several other work about actor–action semantic segmentation. For example, Kalogeiton et al. (2017) introduced an end-to-end multitask objective that jointly learned object–action relationships and compared with different training objectives. Gavrilyuk et al. (2018) proposed a fully-convolutional model for pixel-level actor and action segmentation using an encoder–decoder architecture optimized for video. They inferred the segmentation from a natural language input sentence. Dang et al. (2018) proposed an end-to-end region-based actor–action segmentation approach which relied on region masks from an instance segmentation algorithm. Compared with our proposed method, Dang et al. (2018) is a supervised approach rather than a weakly-supervised approach which means more supervision is needed using their proposed semantic proposals approach. Moreover, as indicated in Dang et al. (2018), to generate accurate region masks, the method needs fully convolution instance segmentation (FCIS) model trained on specific A2D dataset rather than more generic COCO dataset. Otherwise, too much irrelevant background region will appear in the final results which significantly harm the actor–action segmentation performance. This actually prevents their method to be used in practical since they need FCIS model trained on the specific dataset. However, there is no these requirements for our proposed method.

Our work is also related to many works in semantic video segmentation. Liu and He (2015) proposed an object-augmented dense CRF in the spatio-temporal domain, which captured long-range dependencies between supervoxels and imposed consistency between object and supervoxel labels for multiclass video semantic segmentation. Kundu et al. (2016) extended the fully connected CRF (Krähenbühl and Koltun 2011b) to work on videos. Ladicky et al. (2014) built a hierarchical CRF on multi-scale segmentations that leveraged higher-order potentials in inference. Despite the lack of explicit consideration of actors and actions, we compare to a representative subset of these methods (Krähenbühl and Koltun 2011b; Ladicky et al. 2014) in Sect. 5.

There are many weakly supervised video segmentation methods (Zhong et al. 2016; Zhang et al. 2015, 2017; Liu et al. 2014; Tang et al. 2013; Hartmann et al. 2012) and co-segmentation methods (Tsai et al. 2016; Fu et al. 2014; Wang et al. 2014; Zhang et al. 2014; Chen and Fritz 2013). Zhong et al. (2016) proposed a scene co-parsing framework to assign semantic pixel-wise labels in weakly-labeled videos. Zhang et al. (2017) proposed a novel self-paced fine-tuning network (SPFTN)-based framework, which can learn to explore the context information within the video frames and capture adequate object semantics without using the negative videos. Zhang et al. (2015) proposed a segmentation-by-detection framework to segment objects with video-level tags. Chen and Fritz (2013) studied 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. (2016) proposed 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 lacked 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 (Tsai et al. 2016) (see Sect. 5).

2.5 Multi-task Learning and Ranking

Multi-task learning (MTL) is effective in many applications, such as object detection (Salakhutdinov et al. 2011) and classification (Luo et al. 2013; Yan et al. 2013, 2014, 2016). 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 Argyriou et al. (2007). 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 Jalali et al. (2010) with the goal of identifying irrelevant (outlier) 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 (Jacob et al. 2008; Zhang and Yeung 2010; Zhou et al. 2011a). Different from previous multi-task classification and regression problems, we propose a Schatten p-norm robust multi-task ranking model with the ability to identify outlier tasks. Meanwhile, an efficient solver is devised in this paper.

Ranking SVM is a typical method of learning to rank and has been widely used in information retrieval (CAO et al. 2006). Learning to rank can be categorized into point-wise, pair-wise and list-wise approaches. In point-wise methods, the higher ranked items are assigned higher target scores. Pair-wise methods capture some structure by posing the task as a classification problem over all pairs. List-wise methods wrestle with the full combinatorial structure and thus have to deal with formidable optimization problems. Sculley (2010) proposed using stochastic gradient descent to optimize a linear combination of a pointwise quadratic loss and a pairwise hinge loss from ranking SVM. Amini et al. (2008) presented a boosting based algorithm for learning a bipartite ranking function with partially labeled data. Different from existing ranking methods, we extended ranking SVM to a multi-task setting and provided an efficient solver.

3 Schatten *p*-Norm 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 (Liu 2009). To obtain good potentials for segmentation and select representative supervoxels and action tubes for specific categories (details in Sect. 4), we propose a Schatten p-norm 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 Schatten p-norm robust multi-task ranking.

3.1 Ranking SVM

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.

3.2 Robust Multi-task Ranking

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 (Argyriou et al. 2007) and their dependencies (Jacob et al. 2008; Zhang and Yeung 2010; Zhou et al. 2011a) 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. (2011) 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 $\|\mathbf{L}\|_* = \text{trace}(\sqrt{\mathbf{L}^*\mathbf{L}})$ 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 Chen et al. (2011), our proposed optimization is different in three critical aspects: (i) the optimization problem in Chen et al. (2011) is a regression problem while ours is a ranking optimization problem. This makes (Chen et al. 2011) 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 Chen et al. (2011) 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 (Chen et al. 2011) and our problem.

3.3 Schatten *p*-Norm Robust Multi-task Ranking

Although the trace norm in Eq. 3 is a convex problem, the relaxation may make the solution seriously deviate from the original solution. It is desired to solve a better approximation of the rank minimization problem without introducing

much computational cost. We reformulate the robust multi-task ranking problem using the Schatten p-norm.

The Schatten p-norm $(0 < \mathrm{p} < \infty)$ of a matrix $\mathbf{A} \in I\!R^{l\times m}$ is defined as

$$\|\mathbf{A}\|_{S_p} = \left(\sum_{i=1}^{min\{l,m\}} \sigma_i^p\right)^{1/p} = \left(tr(\mathbf{A}^T\mathbf{A})^{p/2}\right)^{1/p} \tag{4}$$

where σ_i is the i-th singular value of $\mathbf{A}$ and tr$(\cdot)$ means the trace operator.

The Schatten p-norm of matrix $\mathbf{A} \in I\!R^{l\times m}$ to the power p is

$$\|\mathbf{A}\|_{S_p}^p = \sum_{i=1}^{min\{l,m\}} \sigma_i^p = tr\left(\mathbf{A}^T\mathbf{A}\right)^{p/2} \tag{5}$$

while $p = 1$, the Schatten p-norm becomes trace norm that denoted by $\|\cdot\|_*$ and while $p = 2$, the Schatten p-norm becomes Frobenius norm that denoted by $\|\cdot\|_F$.

Based on the above definition, we extend our robust multi-task ranking with the Schatten p-norm version. The optimization problem becomes

$$\begin{aligned}
&\min_{\mathbf{W},\gamma,\varepsilon} \frac{1}{2}\|\mathbf{W}\|_F^2 + C_1\sum_{i,j\in S}\gamma_{ijk} + C_2\sum_{i,j\in D}\varepsilon_{ijk} \\
&\quad + \lambda_1\|\mathbf{L}\|_{S_p}^p + \lambda_2\|\mathbf{E}\|_{1,2} \\
&s.t.\ \left|\mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk}\right| \le \gamma_{ijk} \\
&\quad \mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk} \ge 1 - \varepsilon_{ijk} \\
&\quad \varepsilon_{ijk} \ge 0 \\
&\quad \gamma_{ijk} \ge 0 \\
&\quad \mathbf{W} = \mathbf{L} + \mathbf{E}.
\end{aligned} \tag{6}$$

3.4 Optimization

The proposed optimization problem in Eq. 6 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. $\mathbf{S}$ is used to replace the explicit decomposition of $\mathbf{W}$ in Eq. 6. Then the mixture of norms can be placed on $\mathbf{S}$ which suggests an update independent from $\mathbf{W}$. Thus, the optimization can be facilitated. The optimization problem can be decomposed into two separate steps by iteratively updating $\mathbf{W}$ and $\mathbf{S}$ respectively. We adopt Proximal Operator Computation approach (Parikh and Boyd 2013). The benefit is that the column vectors of W can be optimized separately. Specifically, each vector of the optimal W can be obtained via solving sub-problems. With the slack variable, the optimization problem becomes,

$$\begin{aligned}
&\min_{\mathbf{W},\mathbf{S},\gamma,\varepsilon} \frac{1}{2}\|\mathbf{W}\|_F^2 + C_1\sum_{i,j\in S}\gamma_{ijk} + C_2\sum_{i,j\in D}\varepsilon_{ijk} \\
&\quad + \|\mathbf{W}-\mathbf{S}\|_F^2 + \lambda\Phi(\mathbf{S}) \\
&s.t.\ \left|\mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk}\right| \le \gamma_{ijk} \\
&\quad \mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk} \ge 1 - \varepsilon_{ijk} \\
&\quad \varepsilon_{ijk} \ge 0 \\
&\quad \gamma_{ijk} \ge 0
\end{aligned} \tag{7}$$

The term $\|\mathbf{W}-\mathbf{S}\|_F^2$ in Eq. 7 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. 7 as follows:

Step 1 Fix $\mathbf{S}$, optimize $\mathbf{W}$. Equation 3 becomes,

$$\begin{aligned}
&\min_{\mathbf{w}_k,\gamma,\varepsilon} \frac{1}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|^2 + C_1\sum_{i,j\in S}\gamma_{ijk} + C_2\sum_{i,j\in D}\varepsilon_{ijk} \\
&\quad + \sum_{k=1}^{K}\|\mathbf{w}_k - \mathbf{s}_k\|^2 \\
&s.t.\ \left|\mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk}\right| \le \gamma_{ijk} \\
&\quad \mathbf{w}_k^T\mathbf{x}_{ik} - \mathbf{w}_k^T\mathbf{x}_{jk} \ge 1 - \varepsilon_{ijk} \\
&\quad \varepsilon_{ijk} \ge 0 \\
&\quad \gamma_{ijk} \ge 0
\end{aligned} \tag{8}$$

Equation 8 can be decomposed into K separate single-task SVM ranking sub-problems and therefore can be solved via a standard SVM ranking solver (Joachims 2006).

Step 2 Fix $\mathbf{W}$, optimize $\mathbf{S}$. Equation 3 becomes,

$$\min_{\mathbf{S}} \|\mathbf{S}-\mathbf{W}\|_F^2 + \lambda\Phi(\mathbf{S}) \tag{9}$$

The first term in Eq. 9 penalizes the learned slack weight matrix $\mathbf{S}$ to be close to the original matrix $\mathbf{W}$. $\Phi(\mathbf{S})$ can be $\|\mathbf{S}\|_{S_p}^p$. Solving the problem Eq. 9 is challenge since the nonsmooth and intractable of Schatten p-norm. We use the augmented Lagrangian multiplier (ALM) method (Dp 1996) to solve this problem.

The Eq. 9 can be equivalently rewritten as

$$\min_{\mathbf{S},\mathbf{P}=\mathbf{S}-\mathbf{W},\mathbf{S}=\mathbf{Z}} \|\mathbf{P}\|_F^2 + \gamma\|\mathbf{Z}\|_{S_p}^p \tag{10}$$

Based on Augmented Lagrangian Multiplier method, we solve the following problem:

$$\begin{aligned}
&\min_{\mathbf{S},\mathbf{P},\mathbf{Z}} \|\mathbf{P}\|_F^2 + \gamma\|\mathbf{Z}\|_{S_p}^p + \frac{\mu}{2}\left\|\mathbf{P} - (\mathbf{S}-\mathbf{W}) + \frac{1}{\mu}\mathbf{\Lambda}\right\|_F^2 \\
&\quad + \frac{\mu}{2}\left\|\mathbf{S} - \mathbf{Z} + \frac{1}{\mu}\mathbf{\Sigma}\right\|_F^2
\end{aligned} \tag{11}$$

We use alternating direction method (ADM) (Gabay and Mercier 1976) to solve the problem with respect to $\mathbf{S}$, $\mathbf{P}$, $\mathbf{Z}$.

(i) While fixing $\mathbf{P}$, $\mathbf{Z}$, the problem (11) is simplified to the following problem:

$$\min_{\mathbf{S}} \|\mathbf{S}-\mathbf{Q}\|_F^2 + \|\mathbf{S}-\mathbf{R}\|_F^2 \tag{12}$$

where $\mathbf{Q} = \mathbf{P} + \mathbf{W} + \frac{1}{\mu}\mathbf{\Lambda}$ and $\mathbf{R} = \mathbf{Z} - \frac{1}{\mu}\mathbf{\Sigma}$. The optimal solution to problem (12) can be easily obtained by $\mathbf{S} = (\mathbf{Q}+\mathbf{R})/2$.

(ii) While fixing $\mathbf{S}$, $\mathbf{Z}$, the problem (11) is simplified to the following problem:

$$\min_{\mathbf{P}} \|\mathbf{P}\|_F^2 + \frac{\mu}{2}\|\mathbf{P}-\mathbf{H}\|_F^2 \tag{13}$$

where $\mathbf{H} = \mathbf{S} - \mathbf{W} - \frac{1}{\mu}\mathbf{\Lambda}$, the optimal solution $\mathbf{P} = \frac{\mu}{2+\mu}\mathbf{H}$.

(iii) While fixing $\mathbf{S}$, $\mathbf{P}$, the problem (11) is simplified to the following problem:

$$\min_{\mathbf{Z}} \gamma \|\mathbf{Z}\|_{S_p}^p + \frac{\mu}{2}\|\mathbf{Z}-\mathbf{B}\|_F^2 \tag{14}$$

where $\mathbf{B} = \mathbf{S} + \frac{1}{\mu}\mathbf{\Sigma}$. The optimal solution for $\mathbf{Z}$ is $\mathbf{U1V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vector matrices of $\mathbf{B}$, respectively, and the i-th diagonal element δ_i of the diagonal matrix $\mathbf{\Delta}$.

The algorithm solving the proposed problem is summarized as in Algorithm 1.

Algorithm 1 Solving Eq. 7

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. 8 using Joachims (2006), update $\mathbf{w}_k$
  end
2. Fix $\mathbf{W}$, optimize $\mathbf{S}$
  Optimize Eq. 9 using Augmented Lagrangian Multiplier (ALM)
  (i) Fix $\mathbf{P}$,$\mathbf{Z}$, update $\mathbf{S}$ with $(\mathbf{Q}+\mathbf{R})/\mathbf{2}$
  (ii) Fix $\mathbf{S}$,$\mathbf{Z}$, update $\mathbf{P}$ with $\mu/(2+\mu)\mathbf{H}$
  (iii) Fix $\mathbf{S}$,$\mathbf{P}$, update $\mathbf{Z}$ with $\mathbf{U}\Delta\mathbf{V}$
Until Convergence
Output: W

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) (Grundmann et al. 2010). 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 Sect. 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 Sect. 4.3). Finally, we devise a CRF model for actor–action segmentation (see Sect. 4.4).

4.2 Supervoxels and Action Tubes

4.2.1 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 Xu and Corso (2016b). Based on their work, we adopt the GBH supervoxel segmentation and consider supervoxels from three different levels in a hierarchy. The performance of different levels are evaluated in Sect. 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 (Girshick et al. 2016). The final feature vector representing the *actor* of a superpvoxel is averaged over the three time-slices as shown in Fig. 2b.

4.2.2 Tubes

Each supervoxel defines an action tube that is the sequence of minimum bounding rectangles around the supervoxel over time. Jain et al. (2014) use such action tubes to localize human actions in videos. Here, we use them as proposals for

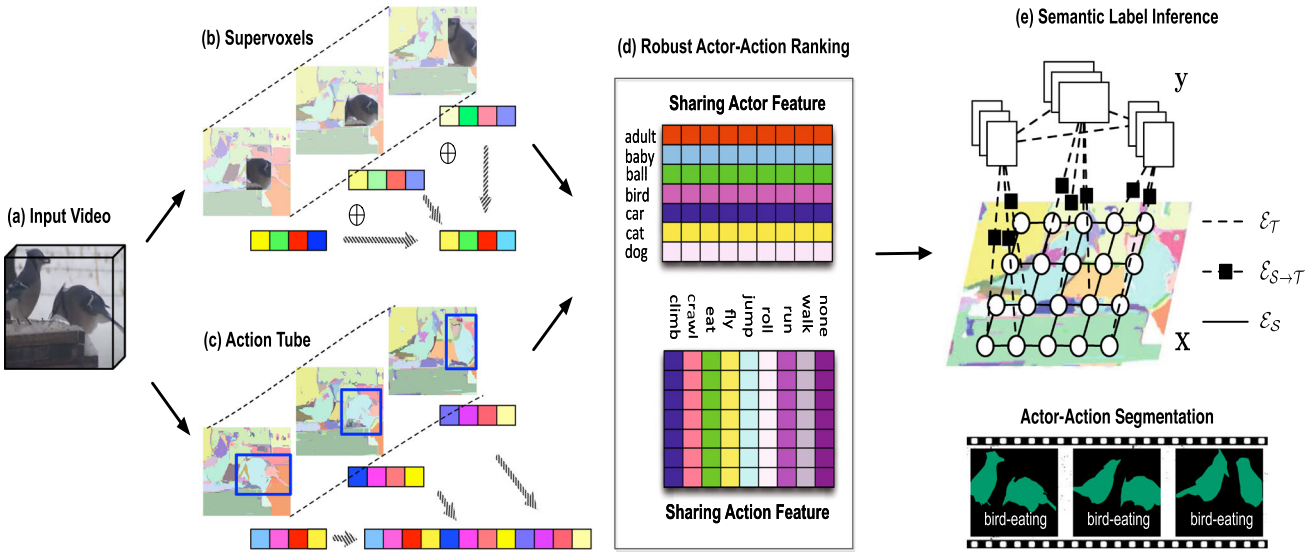

Fig. 2 Overview of our proposed weakly supervised actor–action segmentation framework. **a** Input videos from the video dataset. **b** Supervoxel generation and feature extraction. **c** Action tube generation and feature extraction. **d** Sharing features among different actors and actions. **e** Semantic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

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. 2c.

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 Sect. 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:

$$
\begin{aligned}
(\mathbf{x}^*, \mathbf{y}^*) = \arg\min_{x,y} & \sum_{(i,j)\in\mathcal{E}_{\mathcal{S}}} \psi(x_i, x_j) + \sum_{(i,j)\in\mathcal{E}_{\mathcal{T}}} \psi(y_i, y_j) \\
& + \sum_{i\in\mathcal{S}} \phi(x_i) + \sum_{i\in\mathcal{T}} \varphi(y_i) \\
& + \sum_{(i,j)\in\mathcal{E}_{\mathcal{S}\to\mathcal{T}}} \xi(x_i, y_j) \ , \qquad (15)
\end{aligned}
$$

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 (Xu and Corso 2016a), we also use video-level potentials as an additional global labeling cost. Comparing to the models in Xu et al. (2015), our model is more flexible and allows separate topologies for supervoxels and action tubes (see Fig. 2e).

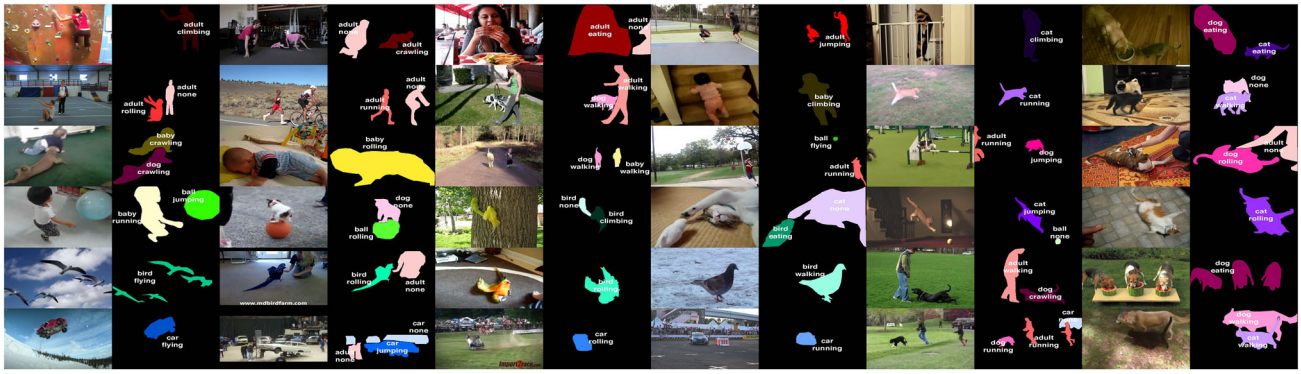

Fig. 3 Examples from actor–action (A2D) video dataset

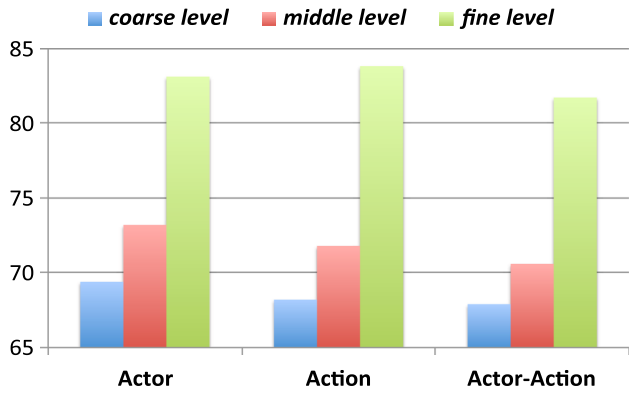

Fig. 4 The overall pixel accuracy for different GBH hierarchy supervoxels on A2D dataset

Finally the segmentation is generated by mapping action tubes to supervoxels.

CRF models are the most effective approaches for image and video segmentation (Fulkerson et al. 2009). Basic CRF models are composed of unary potentials on individual pixels/voxels or superpixels/supervoxels, and pairwise potentials on neighboring pixels/voxels or superpixels/supervoxels. Inspired by Xu et al. (2015), we represent actor nodes and action nodes as two separate CRF layers to perform actor–action semantic segmentation. Bi-layer CRF model connects each pair of random variables with an edge encodes the potentials. The unary and pair-wise potentials are learned via proposed multi-task ranking approach.

5 Experiments

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

5.1 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 (Xu et al. 2015) as shown in Fig. 3, 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.

Another dataset used in the experiments is Youtube-objects dataset (Prest et al. 2012) which contains 10 object categories, e.g. *aeroplane, bird, boat, car, cat, cow, dog, horse, motorbike, train*, and the length of each sequence is up to 400 frames. Since there are no action labels for videos in the Youtube-objects dataset, we extend the dataset for actor–action analysis by adding action labels to videos, e.g. *climbing, crawling, eating, flying, jumping, rolling, running, walking*. We evaluate the proposed algorithm in a subset of 126 videos with more than 20,000 frames, where the pixel-wise annotations in every 10 frames are provided by Jain and Grauman (2014).

5.2 Experimental Settings

We use GBH (Grundmann et al. 2010) 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 to 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 (Szegedy et al. 2015) 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. Parameter p in Schatten p-norm is grid-searched via range $[0.1, 0.2, \ldots, 0.9, 1]$ in the experimental setting. 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 (Delong et al. 2012) for CRF inference and empirically set the parameters by hand. We follow the same setup as Xu et al. (2015) for the training/testing split of the dataset.

5.3 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.

5.4 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. 4 for A2D dataset and Fig. 5 for Youtube-objects dataset, respectively. 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 super-

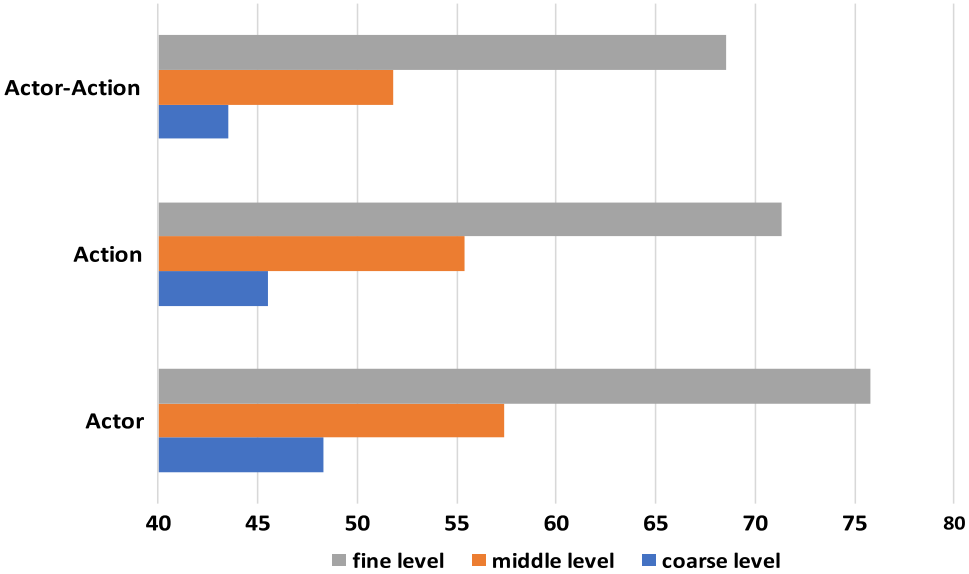

Fig. 5 The overall pixel accuracy for different GBH hierarchy supervoxels on Youtube-objects dataset

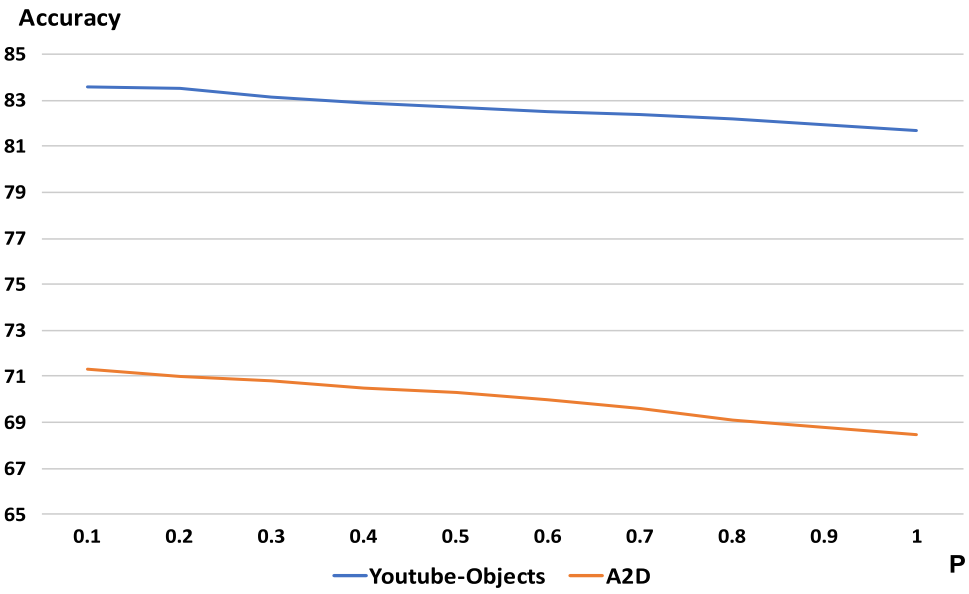

Fig. 6 The overall pixel accuracy for different value of p on both A2D and Youtube-objects dataset

voxels (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 overall pixel accuracy of 83.6% on A2D dataset and 71.3% on Youtube-objects dataset when we use both coarse labels (actor and action) and fine-grained labels (actor–action). Meanwhile, we only achieve overall pixel accuracy of 72.6% on A2D dataset and 57.4% on Youtube-objects dataset 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.

We evaluate the performance w.r.t different p values in our Schatten p-norm robust multi-task ranking framework. We vary the value of p in the range of $\{0.1,\ 0.2, \ldots, 1\}$. Fig. 6 shows the performance of overall pixel accuracy for A2D and Youtube-objects datasets. We observe that the overall pixel accuracy increase when the value of p decreases. This result clearly justifies the effectiveness of the proposed Schatten p-norm in the proposed robust multi-task ranking approach.

Table 1 Comparison of overall pixel accuracy on the A2D dataset (the top pixel-level, frame-level and video-level results are high-lighted)

	Action	Actor	Actor–action	Label
AHRF (Ladicky et al. 2014)	63.9	64.9	63.0	Pixel-level
GPM (Xu and Corso 2016a)	82.4	82.2	80.8	Pixel-level
FCRF (Krähenbühl and Keltun 2011a)	77.6	77.9	76.2	Pixel-level
JSS (Ji et al. 2018)	92.6	94.5	92.5	Pixel-level
Pixel-level	**92.6**	**94.5**	**92.5**	–
RM (Dang et al. 2018)	93.4	95.3	93.0	Frame-level
Frame-level	**93.4**	**95.3**	**93.0**	–
RSVM (Joachims 2006)	70.1	70.8	68.8	Video-level
DM-MTL (Jalali et al. 2010)	72.3	72.9	71.4	Video-level
C-MTL (Zhou et al. 2011a)	73.1	73.5	72.7	Video-level
MT-Lasso (Tibshirani 1996)	67.3	68.1	65.2	Video-level
MR-MTL (Evgeniou and Pontil 2004)	68.1	68.6	66.7	Video-level
WSS (Tsai et al. 2016)	71.5	71.9	70.4	Video-level
Ours ($p = 1$) (Yan et al. 2017)	83.8	83.1	81.7	Video-level
Ours ($p = 0.1$)	85.7	86.2	83.6	Video-level
Video-level	**85.7**	**86.2**	**83.6**	–

Table 2 Comparison of overall pixel accuracy on the Youtube-objects dataset (the top pixel-level, frame-level and video-level results are high-lighted)

	Action	Actor	Actor–action	Label
AHRF (Ladicky et al. 2014)	62.6	63.7	60.1	Pixel-level
GPM (Xu and Corso 2016a)	70.1	73.2	66.8	Pixel-level
FCRF (Krähenbühl and Keltun 2011a)	64.6	64.9	59.2	Pixel-level
JSS (Ji et al. 2018)	74.6	78.1	72.5	Pixel-level
Pixel-level	**74.6**	**78.1**	**72.5**	–
RM (Dang et al. 2018)	75.9	78.5	73.6	Frame-level
Frame-level	**75.9**	**78.5**	**73.6**	–
RSVM (Joachims 2006)	56.1	57.2	51.5	Video-level
DM-MTL (Jalali et al. 2010)	69.3	68.9	63.1	Video-level
C-MTL (Zhou et al. 2011a)	69.5	69.3	63.4	Video-level
MT-Lasso (Tibshirani 1996)	62.3	63.1	60.7	Video-level
MR-MTL (Evgeniou and Pontil 2004)	63.3	64.2	61.4	Video-level
WSS (Tsai et al. 2016)	64.5	65.3	61.3	Video-level
Ours ($p = 1$) (Yan et al. 2017)	71.3	75.8	68.5	Video-level
Ours ($p = 0.1$)	73.1	77.4	71.3	Video-level
Video-level	**73.1**	**77.4**	**71.3**	–

Table 3 Comparison of per-class accuracy on the A2D dataset (top-2 scores for each category are highlighted)

Method	BK	Baby					Ball				Car				
		Climb	Crawl	Roll	Walk	None	Fly	Jump	Roll	None	Fly	Jump	Roll	Run	None
AHRF (Ladicky et al. 2014)	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 (Xu and Corso 2016a)	**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 (Krähenbühl and Keltun 2011a)	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
JSS (Ji et al. 2018)	**87.5**	**70.8**	72.1	57.3	**64.2**	7.1	14.5	**25.7**	24.8	5.2	43.1	**82.5**	**71.2**	**61.8**	0.0
RSVM (Joachims 2006)	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 (Jalali et al. 2010)	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 (Zhou et al. 2011a)	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
MT-Lasso (Tibshirani 1996)	75.3	48.2	45.6	50.2	41.3	0.0	4.4	8.6	12.1	0.0	25.8	54.5	34.1	38.1	0.0
MR-MTL (Evgeniou and Pontil 2004)	77.1	48.5	48.1	51.3	43.1	0.0	5.5	9.7	14.2	0.0	28.2	57.5	35.1	39.2	0.0
WSS (Tsai et al. 2016)	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 ($p = 1$) (Yan et al. 2017)	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**
Ours ($p = 0.1$)	84.0	**68.4**	**76.3**	**79.8**	54.8	**36.0**	**21.3**	22.4	**65.1**	**17.1**	**48.3**	67.4	44.2	51.6	**24.9**

Method	Adult								Bird						
	Climb	Crawl	Eat	Jump	Roll	Run	Walk	None	Climb	Eat	Fly	Jump	Roll	Walk	None
AHRF (Ladicky et al. 2014)	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 (Xu and Corso 2016a)	**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 (Krähenbühl and Keltun 2011a)	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
JSS (Ji et al. 2018)	**87.3**	**85.0**	**79.2**	**51.4**	49.2	56.5	45.3	**78.0**	**57.8**	32.1	**65.3**	**19.7**	**49.7**	**51.2**	**4.3**
RSVM (Joachims 2006)	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 (Jalali et al. 2010)	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

Table 3 continued

Method	Adult								Bird						
	Climb	Crawl	Eat	Jump	Roll	Run	Walk	None	Climb	Eat	Fly	Jump	Roll	Walk	None
C-MTL (Zhou et al. 2011a)	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**
MT-Lasso (Tibshirani 1996)	34.2	35.6	55.2	20.4	28.3	28.7	24.4	0.0	40.1	19.5	42.3	10.2	25.9	21.6	0.0
MR-MTL (Evgeniou and Pontil 2004)	36.5	26.4	56.1	21.3	29.1	28.9	24.7	0.0	41.1	19.8	43.3	11.4	26.2	23.2	0.0
WSS (Tsai et al. 2016)	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 ($p = 1$) (Yan et al. 2017)	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
Ours ($p = 0.1$)	47.7	49.9	77.3	35.9	**52.3**	45.0	**48.4**	**56.2**	49.5	**33.9**	53.5	15.9	35.0	33.3	0.0

Method	Dog							Cat							Avg
	Crawl	Eat	Jump	Roll	Run	Walk	None	Climb	Eat	Jump	Roll	Run	Walk	None	–
AHRF (Ladicky et al. 2014)	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 (Xu and Corso 2016a)	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 (Krähenbühl and Keltun 2011a)	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
JSS (Ji et al. 2018)	47.3	65.2	35.2	64.6	22.1	**71.5**	**54.2**	**49.2**	59.2	**37.7**	81.9	**59.0**	39.1	**75.2**	**51.4**
RSVM (Joachims 2006)	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 (Jalali et al. 2010)	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 (Zhou et al. 2011a)	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
MT-Lasso (Tibshirani 1996)	28.2	58.3	18.8	41.2	14.6	50.1	0.0	10.1	38.3	7.3	57.4	17.3	4.5	0.0	26.6
MR-MTL (Evgeniou and Pontil 2004)	29.9	59.4	19.0	41.8	15.1	51.2	0.0	11.2	39.2	9.1	58.3	18.6	5.7	0.0	27.4
WSS (Tsai et al. 2016)	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 ($p = 1$) (Yan et al. 2017)	**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
Ours ($p = 0.1$)	**67.1**	**87.5**	**53.2**	**75.3**	**69.8**	64.3	**9.4**	**43.2**	**74.3**	**38.6**	**88.5**	39.3	**67.5**	**28.1**	**44.9**

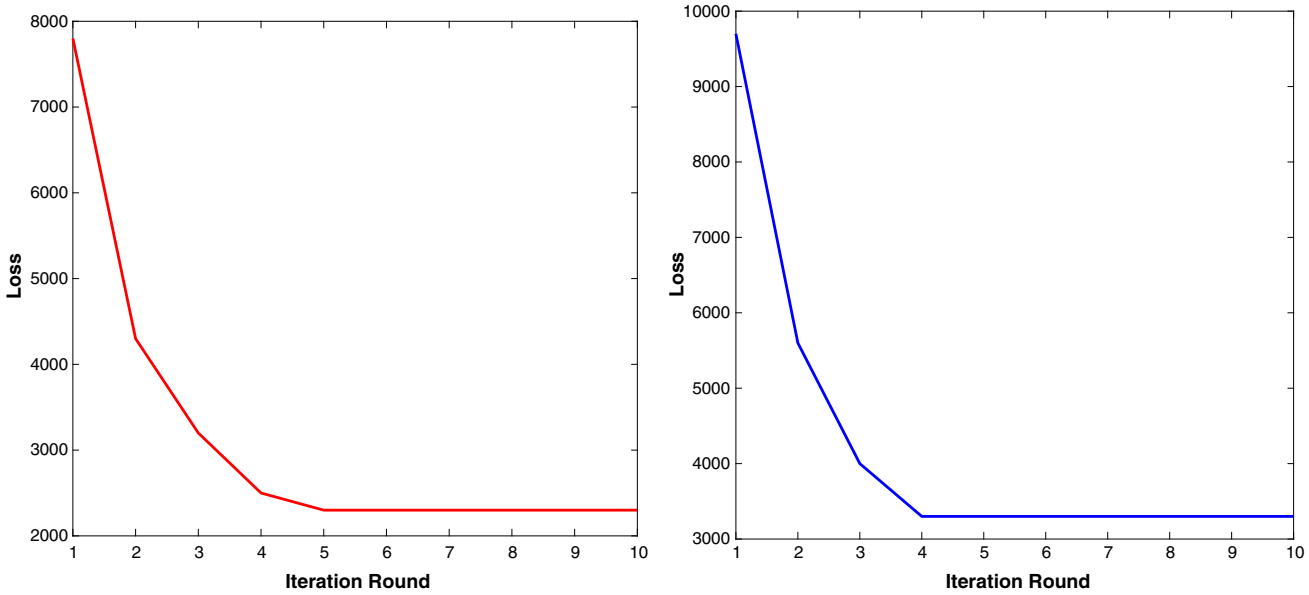

Fig. 7 Convergence of schatten p-norm robust multi-task ranking algorithm on (left) A2D dataset and (right) Youtube-object dataset

5.5 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) (Ladicky et al. 2014), Grouping Process Models (GPM) (Xu and Corso 2016a), fully-connected CRF (FCRF) (Krähenbühl and Keltun 2011a), Region Mask (RM) (Dang et al. 2018) and Joint Semantic Segmentation (JSS) (Ji et al. 2018). 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) (Tsai et al. 2016). 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), Multi-task Lasso (MT-Lasso) (Tibshirani 1996), mean-regularized multi-task learning (MR-MTL) (Evgeniou and Pontil 2004), dirty model multi-task learning (DM-MTL) (Jalali et al. 2010), and clustered multi-task learning (C-MTL) (Zhou et al. 2011a). For fair comparison, we use author-released code for methods (Xu and Corso 2016a; Tsai et al. 2016). For Ranking SVM, we use the released implementation in Joachims (2006). For multi-task learning approaches (Jalali et al. 2010; Zhou et al. 2011a; Tibshirani 1996; Evgeniou and Pontil 2004), we use the MALSAR toolbox (Zhou et al. 2011b). 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.

Tables 1 and 2 show the overall pixel accuracy for all methods on A2D and Youtube-objects datasets respectively. We observe that our method outperforms all other baselines except JSS (Ji et al. 2018) and RM (Dang et al. 2018). However, we note that JSS is a fully supervised approach, i.e. it needs tedious pixel-level human labelling for training samples. We performed additional experiments on adopting semantic proposals as in Dang et al. (2018) in the experimental section. As we observed from Tables 1 and 2, there is 9% and 2% performance increasing on A2D and Youtube-objects datasets respectively by adopting semantic proposals. However, there are additional costs for the semantic proposals approach. First, this is a supervised approach rather than a weakly-supervised approach. This means more supervision is needed using the semantic proposals approach. Second, as indicated in Dang et al. (2018), to generate accurate region masks, the method needs fully convolution instance segmentation (FCIS) model trained on specific A2D dataset rather than more generic COCO dataset. Otherwise, too much irrelevant background region will appear in the final results which significantly harm the actor–action segmentation performance (3% and 8% lower than our approach on A2D and Youtube-objects datasets). This actually prevents their method to be used in practical since they need FCIS model trained on the specific dataset. Our approach has 13%/10% higher accuracy than the other weakly supervised approach (WSS) (Tsai et al. 2016) on A2D/Youtube-objects datasets. 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, MT-Lasso, MR-MTL) approaches by up to 15%, 12%, 11%, 18%, 17% (A2D dataset) and 20%, 8%, 8%, 11%, 10% (Youtube-objects

Fig. 8 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 (Xu and Corso 2016a), WSS (Tsai et al. 2016), RSVM (Joachims 2006), DM-MTL (Jalali et al. 2010) and AHRF (Ladicky et al. 2014) respectively. Our method is able to generate correct actor–action segmentation expect for cat-jumping and adult-running in these examples

dataset) respectively, which shows the robustness of our approach.

Table 3 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 JSS (Ji et al. 2018). However, we note that JSS 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.

We also analyze the convergence rate and computational cost for our proposed Schatten p-norm Robust Multi-task Ranking approach. The proposed iterative approach monotonically decreases the objective function value in Eq. 7 until convergence. Figure 7 shows the convergence curves of our algorithm on A2D dataset and Youtube-objects dataset. It can be observed that the objective function value converges quickly and our approach usually converges after 5 iterations at most (precision $= 10^{-5}$). Regarding the computational cost of our proposed algorithm, we train our model on A2D dataset in 9 min without cross-validation on a workstation with Intel Core i7 (8th Gen) i7-8700K 3.70 GHz CPU processor and NVIDIA GeForce GTX 1080 Ti GPU. This means our algorithm would be scalable for large-scale video problems. We also compare our Schatten p-norm Robust Multi-task Ranking approach with Yan et al. (2017), where we train the model on A2D dataset in 8 min without cross-validation. Since the more advanced alternating direction optimization method adopted, the computation cost of our proposed Schatten p-norm version is in the same computational level as Yan et al. (2017).

Figure 8 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 simultaneously and motion is significant in the video.

6 Conclusion

In conclusion, modeling and generating realistic human behavior data is an important research topic in literature. Fine-grained activity understanding in videos is a key step to achieve this goal. In this paper, we propose a novel weakly supervised actor–action segmentation method. Particularly, a Schatten p-norm 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 both the large-scale A2D dataset and Youtube-objects dataset show the effectiveness of our proposed approach. Our approach is able to generate fine-grained actor–action video semantic segmentation maps which can be further used for behavior understanding. After segmentation of actors in video sequences, the next step is to recognize and understand the behaviors of actors. The essence of behavior understanding may be considered to be a classification problem towards time varying data.

Acknowledgements This research was partially supported by a University of Michigan MiBrain Grant (DC, JC), DARPA FA8750-17-2-0112 (JC), National Institute of Standards and Technology Grant 60NANB17D191 (JC, YY), NSF IIS-1741472 and IIS-1813709 (CX), NSF NeTS-1909185 and CSR-1908658 (YY), and gift donation from Cisco Inc (YY). This article solely reflects the opinions and conclusions of its authors and not the funding agents.

References

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S. (2016). *Youtube-8m: A large-scale video classification benchmark*. Technical report. Preprint arXiv:1609.08675.

Amini, M. R., Truong, T. V., & Goutte, C. (2008). A boosting algorithm for learning bipartite ranking functions with partially labeled data. In *SIGIR*.

Argyriou, A., Evgeniou, T., & Pontil, M. (2007). Multi-task feature learning. In *NIPS*.

Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., & Sivic, J. (2014). Weakly supervised action labeling in videos under ordering constraints. In *ECCV*.

Brendel, W., & Todorovic, S. (2009). Video object segmentation by tracking regions. In *ICCV*.

Brox, T., & Malik, J. (2010). Object segmentation by long term analysis of point trajectories. In *ECCV*.

Caba Heilbron, F., Escorcia, V., Ghanem, B., & Carlos Niebles, J. (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*.

Cao, Y., Xu, J., Liu, T. Y., Li, H., Huang, Y., & Hon, H. W. (2006). Adapting ranking SVM to document retrieval. In *SIGIR*.

Chao, Y. W., Wang, Z., Mihalcea, R., & Deng, J. (2015). Mining semantic affordances of visual object categories. In *CVPR*.

Chen, J., Zhou, J., & Ye, J. (2011). Integrating low-rank and group-sparse structures for robust multi-task learning. In *ACM SIGKDD conferences on knowledge discovery and data mining*.

Chen, W., & Corso, J. J. (2015). Action detection by implicit intentional motion clustering. In *ICCV*.

Chiu, W. C., & Fritz, M. (2013). Multi-class video co-segmentation with a generative multi-video model. In *CVPR*.

Corso, J. J., Sharon, E., Dube, S., El-Saden, S., Sinha, U., & Yuille, A. (2008). Efficient multilevel brain tumor segmentation with integrated Bayesian model classification. *IEEE Transactions on Medical Imaging*, *27*, 629–640.

Dang, K., Zhou, C., Tu, Z., Hoy, M., Dauwels, J., & Yuan, J. (2018). Actor action semantic segmentation with region masks. In *BMVC*.

Delong, A., Osokin, A., Isack, H. N., & Boykov, Y. (2012). Fast approximate energy minimization with label costs. *International Journal of Computer Vision*, *96*(1), 1–27.

Deselaers, T., Alexe, B., & Ferrari, V. (2012). Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, *100*(3), 275–293.

Dp, B. (1996). *Constrained optimization and lagrange multiplier methods*. Belmont: Athena Scientific.

Evgeniou, T., & Pontil, M. (2004). Regularized multi-task learning. In *KDD*.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision*, *59*(2), 167–181.

Fu, H., Xu, D., Zhang, B., & Lin, S. (2014). Object-based multiple foreground video co-segmentation. In *CVPR*.

Fulkerson, B., Vedaldi, A., & Soatto, S. (2009). Class segmentation and object localization with superpixel neighborhoods. In *ICCV*.

Gabay, D., & Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Applications*, *2*(1), 17–40.

Galasso, F., Cipolla, R., & Schiele, B. (2012). Video segmentation with superpixels. In *Asian conference on computer vision*.

Gavrilyuk, K., Ghodrati, A., Li, Z., & Snoek, C. G. (2018). Actor and action video segmentation from a sentence. In *CVPR*.

Geest, R. D., Gavves, E., Ghodrati, A., Li, Z., Snoek, C., & Tuytelaars, T. (2016). Online action detection. In *ECCV*.

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

Grundmann, M., Kwatra, V., Han, M., & Essa, I. (2010). Efficient hierarchical graph-based video segmentation. In *CVPR*.

Guo, J., Li, Z., Cheong, L. F., & Zhou, S. Z. (2013). Video co-segmentation for meaningful action extraction. In *ICCV*.

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

Hartmann, G., Grundmann, M., Hoffman, J., Tsai, D., Kwatra, V., Madani, O., et al. (2012). Weakly supervised learning of object segmentations from web-scale video. In *ECCV workshops* (pp. 198–208). Berlin: Springer.

Iwashita, Y., Takamine, A., Kurazume, R., & Ryoo, M. S. (2014). First-person animal activity recognition from egocentric videos. In *IEEE international conference on pattern recognition*.

Jacob, L., Bach, F., & Vert, J. (2008). Clustered multi-task learning: A convex formulation. In *NIPS*.

Jain, M., Van Gemert, J., Jégou, H., Bouthemy, P., & Snoek, C., et al. (2014). Action localization with tubelets from motion. In *CVPR*.

Jain, S., & Grauman, K. (2014). Supervoxel-consistent foreground propagation in video. In *ECCV*.

Jalali, A., Ravikumar, P., Sanghavi, S., & Ruan, C. (2010). A dirty model for multi-task learning. In *NIPS*.

Ji, J., Buch, S., Soto, A., & Niebles, J. C. (2018). End-to-end joint semantic segmentation of actors and actions in video. In *ECCV*.

Joachims, T. (2006). Training linear SVMs in linear time. In *ACM SIGKDD conferences on knowledge discovery and data mining*.

Joulin, A., Tang, K., & Fei-Fei, L. (2014). Efficient image and video co-localization with Frank–Wolfe algorithm. In *ECCV*.

Kalogeiton, V., Weinzaepfel, P., Ferrari, V., & Schmid, C. (2017). Joint learning of object and action detectors. In *ICCV*.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*.

Krähenbühl, P., & Keltun, V. (2011a). Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*.

Krähenbühl, P., & Koltun, V. (2011b). Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*.

Kumar, M., Torr, P., & Zisserman, A. (2005). Learning layered motion segmentations of video. In *ICCV*.

Kundu, A., Vineet, V., & Koltun, V. (2016). Feature space optimization for semantic video segmentation. In *CVPR*.

Ladicky, L., Russell, C., Kohli, P., & Torr, P. (2014). Associative hierarchical random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(6), 1056–1077.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *CVPR*.

Lea, C., Reiter, A., Vidal, R., & Hager, G.D. (2016). Segmental spatiotemporal CNNs for fine-grained action segmentation. In *ECCV*.

Lezama, J., Alahari, K., Josef, S., & Laptev, I. (2011). Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *CVPR*.

Lin, G., Shen, C., van den Hengel, A., & Reid, I. (2016). Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*.

Liu, B., & He, X. (2015). Multiclass semantic video segmentation with object-level active inference. In *CVPR*.

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

Liu, X., Tao, D., Song, M., Ruan, Y., Chen, C., & Bu, J. (2014). Weakly supervised multiclass video segmentation. In *CVPR*.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *CVPR*.

Lu, J., Xu, R., & Corso, J. J. (2015). Human action segmentation with hierarchical supervoxel consistency. In *CVPR*.

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

Mettes, P., van Gemert, J. C., & Snoek, C. G. (2016). Spot on: Action localization from pointly-supervised proposals. In *ECCV*.

Mosabbeb, E. A., Cabral, R., De la Torre, F., & Fathy, M. (2014). Multi-label discriminative weakly-supervised human activity recognition and localization. In *Asian conference on computer vision*.

Parikh, N., & Boyd, S. (2013). Proximal algorithms. *Foundations and Trends® in Optimization*, *1*(3), 127–239.

Paris, S. (2008). Edge-preserving smoothing and mean-shift segmentation of video streams. In *ECCV*.

Peng, X., & Schmid, C. (2016). Multi-region two-stream R-CNN for action detection. In *ECCV*.

Pinto, L., Gandhi, D., Han, Y., Park, Y. L., & Gupta, A. (2016). The curious robot: Learning visual representations via physical interactions. In *ECCV*.

Prest, A., Leistner, C., Civera, J., Schmid, C., & Ferrari, V. (2012). Learning object class detectors from weakly annotated video. In *CVPR*.

Rodriguez, M., Ahmed, J., & Shah, M. (2008). Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*.

Ryoo, M. S., & Aggarwal, J. K. (2009). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*.

Salakhutdinov, R., Torralba, A., & Tenenbaum, J. (2011). Learning to share visual appearance for multiclass object detection. In *CVPR*.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local SVM approach. In *IEEE international conference on pattern recognition*.

Sculley, D. (2010). Combined regression and ranking. In *KDD*.

Shou, Z., Wang, D., & Chang, S. F. (2016). Temporal action localization in untrimmed videos via multi-stage CNNs. In *CVPR*.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *NIPS*.

Song, Y. C., Naim, I., Al Mamun, A., Kulkarni, K., Singla, P., Luo, J., Gildea, D., & Kautz, H. (2016). Unsupervised alignment of actions

in video with text descriptions. In *International joint conference on artificial intelligence*.
Soomro, K., Idrees, H., & Shah, M. (2016). Predicting the where and what of actors and actions through online action localization. In *CVPR*.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *CVPR*.
Tang, K., Joulin, A., Li, L. J., & Fei-Fei, L. (2014). Co-localization in real-world images. In *CVPR*.
Tang, K., Sukthankar, R., Yagnik, J., & Fei-Fei, L. (2013). Discriminative segment annotation in weakly labeled video. In *CVPR*.
Tian, Y., Sukthankar, R., & Shah, M. (2013). Spatiotemporal deformable part models for action detection. In *CVPR*.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267–288.
Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *ICCV*.
Tsai, Y. H., Zhong, G., Yang, M. H. (2016). Semantic co-segmentation in videos. In *ECCV*.
Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In *ICCV*.
Wang, L., Hua, G., Sukthankar, R., Xue, J., & Zheng, N. (2014). Video object discovery and co-segmentation with extremely weak supervision. In *ECCV*.
Xiong, C., & Corso, J. J. (2012). Coaction discovery: Segmentation of common actions across multiple videos. In *ACM international workshop on multimedia data mining*.
Xu, C., & Corso, J. J. (2012). Evaluation of super-voxel methods for early video processing. In *CVPR*.
Xu, C., & Corso, J. J. (2016a). Actor–action semantic segmentation with grouping process models. In *CVPR*.
Xu, C., & Corso, J. J. (2016b). LIBSVX: A supervoxel library and benchmark for early video processing. *International Journal of Computer Vision*, *119*(3), 272–290.
Xu, C., Hsieh, S. H., Xiong, C., & Corso, J. J. (2015). Can humans fly? Action understanding with multiple classes of actors. In *CVPR*.
Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*.
Yan, Y., Ricci, E., Subramanian, R., Lanz, O., & Sebe, N. (2013). No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *ICCV*.
Yan, Y., Ricci, E., Subramanian, R., Liu, G., Lanz, O., & Sebe, N. (2016). A multi-task learning framework for head pose estimation under target motion. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, *38*(6), 1070–1083.
Yan, Y., Ricci, E., Subramanian, R., Liu, G., & Sebe, N. (2014). Multitask linear discriminant analysis for multi-view action recognition. *IEEE Transactions on Image Processing*, *23*(12), 5599–5611.
Yan, Y., Xu, C., Cai, D., & Corso, J. J. (2017). Weakly supervised actor–action segmentation via robust multi-task ranking. In *CVPR*.
Yang, Y., Li, Y., Fermüller, C., & Aloimonos, Y. (2015). Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In *AAAI conference on artificial intelligence*.
Yu, S., Tresp, V., & Yu, K. (2007). Robust multi-task learning with t-processes. In *ICML*.
Yuan, J., Ni, B., Yang, X., & Kassim, A. A. (2016). Temporal action localization with pyramid of score distribution features. In *CVPR*.
Zhang, D., Javed, O., & Shah, M. (2014). Video object co-segmentation by regulated maximum weight cliques. In *ECCV*.
Zhang, D., Yang, L., Meng, D., & Dong Xu, J. H. (2017). Spftn: A self-paced fine-tuning network for segmenting objects in weakly labelled videos. In *CVPR*.
Zhang, Y., Chen, X., Li, J., Wang, C., & Xia, C. (2015). Semantic object segmentation via detection in weakly labeled video. In *CVPR*.
Zhang, Y., & Yeung, D. (2010). A convex formulation for learning task relationships in multi-task learning. In *Uncertainty in artificial intelligence*.
Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., & Torr, P. (2015). Conditional random fields as recurrent neural networks. In *ICCV*.
Zhong, G., Tsai, Y. H., & Yang, M. H. (2016). Weakly-supervised video scene co-parsing. In *ACCV*.
Zhou, J., Chen, J., & Ye, J. (2011a). Clustered multi-task learning via alternating structure optimization. In *NIPS*.
Zhou, J., Chen, J., & Ye, J. (2011b). MALSAR: Multi-tAsk Learning via StructurAl Regularization. Arizona State University. http://www.public.asu.edu/~jye02/Software/MALSAR