Multi-label Discriminative Weakly-Supervised Human Activity Recognition and Localization

Ehsan Adeli Mosabbeb[1], Ricardo Cabral[2], Fernando De la Torre[2], Mahmood Fathy[1]

[1]Iran University of Science and Technology
[2]Robotics Institute, Carnegie Mellon University

Abstract. Activity recognition in video has become increasingly important due to its many applications ranging from in-home elder care, surveillance, human computer interaction to automatic sports commentary. To date, most approaches to video rely on fully supervised settings that require time consuming and error prone manual labeling. Moreover, existing supervised approaches are typically tailored for classification, not detection problems (the spatial and temporal support of the action has to be detected). Recently, weakly-supervised learning (WSL) approaches were able to learn discriminative classifiers while localizing the action in space and/or time using weak labels. However, existing approaches for WSL provide coarse localization in terms of spatial regions or spatio-temporal volumes. Moreover, it is unclear how to extend current approaches to the multi-label case that is common in practical applications. This paper proposes a matrix completion approach to the problem of WSL for multi-label learning for video. Our approach localizes non-rectangular spatio-temporal discriminative regions that are inferred by clustering regions of common texture and motion features. We illustrate how our approach improves existing WSL and supervised learning techniques in three standard databases: Hollywood, UCF sports, and MSR-II.

1 Introduction

The idea of recognizing actions automatically from videos brims with potential. Solving it enables many tasks, including surveillance, human-computer interaction, patient monitoring, and automatic sports analysis. However, understanding actions in a video sequence remains a challenging problem due to several reasons: (1) there is a large variability in imaging conditions, as well as in how different people perform an action; (2) background clutter and motion blur are common; (3) data arising from video is of high dimensionality; (4) obtaining ground truth labels for every individual action in every frame of a video is cumbersome. Previous works have addressed these issues by introducing different features [1, 2], interest region detectors such as space-time volumes [3] or trajectories [4, 5], and using different classifiers [2, 6–10]. While these methods have improved recognition results, they may find correlations from background context and non-activity related regions, which result in a lack of interpretability of what is being learned. This motivates us to explore learning techniques that rely less on error-prone human annotations, and learn instead from captions describing the entire video.

In this paper, we propose a multi-label WSL approach to efficiently recognize activities and pinpoint their spatio-temporal location on unseen videos. Fig. 1 shows examples of our results on different datasets. We first extract spatio-temporal activity parts

Fig. 1: Our multi-label weakly-supervised approach recognizes activities and pinpoints their spatio-temporal location on unseen videos. This figure shows results on UCF Sports, HOHA and MSR-II datasets. Top: A sample frame and the extracted spatio-temporal activity parts. Bottom: Activities recognized and localized by our method.

throughout the video. Then, we recognize the activity/activities present in the video, along with selecting the activity parts associated with each recognized activity.

Weakly-supervised learning (WSL) approaches such as multiple instance learning (MIL) ([7–10]) have eased the problems in labeling by localizing discriminative regions while learning the classifier. Instead of class labels, MIL defines labels for positive and negative bags, each containing several instances. All instances in negative bags are negative, but there is at least one positive instance in each positive bag, and the goal is to localize the positive instances (see Fig. 2(a)). Unfortunately, the MIL paradigm has two major drawbacks: first, it is non-trivial to extend it to multi-label settings [11]; second, it typically leads to multi-pass algorithms that alternate between classification and localization. This is especially cumbersome on videos, due to the high number of degrees of freedom in voxel/cuboid search. The MIL problem gets even harder if several instances have to occur together in a bag to form a positive sample. This is the case of action recognition, since activities are typically defined by a collection of spatio-temporal parts extracted from a video [5, 7, 12, 13]. Thus, in order to provide accurate spatio-temporal localization, activity parts cannot be labeled individually, but rather be selected coherently throughout the entire dataset.

We explore the fact that instances from the same class usually organize themselves into clusters [14–17] and that low-rank matrix completion [38] can exploit low-rank subspaces to find relations between labels and features. Thus, we jointly cluster instances into subspaces (Fig. 2(b)) and label unknown instances consistently with the clustering, while keeping negative bag instances as negative (Fig. 2(c)). We demonstrate the effectiveness of our joint subspace clustering and classification in weakly-supervised multi-label learning for video activity recognition.

2 Related Work

Many researchers have addressed the problem of activity recognition in video sequences by using space-time interest points [1, 20], dense trajectories [5] and discriminative

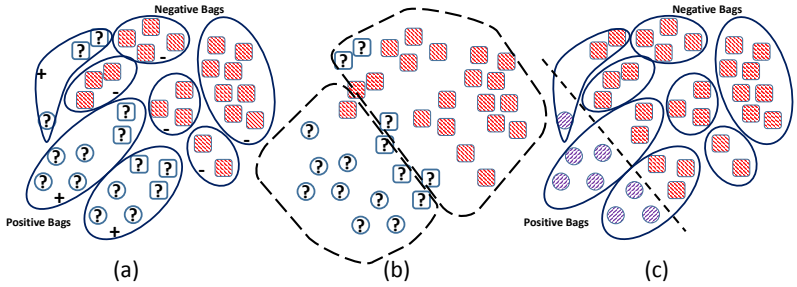


Fig. 2: (a) Multiple instance learning has positive and negative bags, and the goal is to identify positive instances in positive bags. Instead, our approach (b) clusters the instances and (c) forces the labels to agree with the clustering output and bag labels.

space-time neighborhood features [21]. Some previous works have also targeted the problem of spatio-temporal action segmentation and recognition. Hoai *et al*. [22] recognized activities using a multi-class support vector machine (SVM) and infer the temporal segments with dynamic programming. Lan *et al*. [8] trained a latent SVM with a number of labeled and fully annotated videos, but each video is assigned a single label. In [23], the authors propose a weakly supervised video action classification using a similarity constrained latent SVM. Tang *et al*. [24] use a variable-duration hidden Markov model to build a model for each video. Chen *et al*. [25] construct a space-time video graph and find the subgraph that maximizes an activity classifier's score. Siva *et al*. [10] extract potential action cuboids and use genetic algorithms to select the best potential cuboids to learn a SVM for recognition. In related work, [12] introduced spatio-temporal deformable part models for activity recognition and localization.

Action localization is usually performed in the context of action detection, separate from the recognition phase (*e.g*., [26–30]). Raptis *et al*. [7] extract spatio-temporal structures by forming clusters of trajectories. A graphical model is used to recognize a collection of these clusters as a particular action. We share with [7] the use of action parts, but they use graph search to correspond action parts and incorporate fully supervised data, while we perform subspace clustering in a weakly-supervised setting. Ma *et al*. [31] use a two level hierarchical model for activity localization, where each body part is associated with a rectangular box. They first perform a video frame hierarchical segmentation and prune a candidate segment tree. Then they extract hierarchical space-time segments for activity recognition via separate codebooks for root and parts.

Multiple-instance learning was initially proposed in [32] for the WSL problem of predicting which configurations of a pharmaceutical drug are effective. Andrews *et al*. [33] formulated a maximum margin MIL based on Support Vector Machines, where sample labels are unobserved integer variables and the margin between these is maximized directly. These MIL methods result in non-convex optimization processes and thus are heavily dependent on initialization. WSL in computer vision has been extensively studied, by generating spatio-temporal masks for objects in images and videos [34] from partially tagged Internet and YouTube videos [35]. Since labeling video by annotating every single frame is a cumbersome task, several WSL models

have been developed for activity recognition and event detection in videos (*e.g.*, [8, 31]). Tang *et al*. [17] propose a spatio-temporal transductive and inductive object segment annotation from weakly-tagged videos. Recently, several works have formulated the MIL and WSL problems as convex problems (*e.g.*, [36, 37]). In [36] the authors have proposed a model based on calculating likelihood ratios of instances using Support Vector Regression and classifying the bags into positive and negative with a binary SVM.

Our work is most similar to [14] and [38]. [14] is a low-rank subspace segmentation algorithm and [38] a low-rank matrix completion (MC) framework for classification. We propose a method that intertwines these two to perform simultaneous recognition and localization in videos. In [38] each image is represented as a single column in the matrix, localization is performed in the image plane by a bounding-box exhaustive search. However, in our method each video is composed of several parts and supervision is weak and only labels entire videos. Transduction and clustering alone do not suffice, but together provide a selection coherent for all parts in the dataset. This global context means selecting parts yields space-time locations and activity labels.

3 Video Representation

In our method, each video in the dataset is treated as a collection of motion parts [5, 7, 12, 13]. Following [5, 7], videos are represented by features extracted from parts with dense motion trajectories. We perform a spatio-temporal segmentation to obtain volumetric regions that have similar visual and motion characteristics. Then, we extract trajectories using an optical flow tracker, and discard regions with little or no movement. Finally, we group trajectories with similar behavior into parts. Fig. 3 illustrates this process in a sample video from the HOHA dataset. Since trajectories are asynchronous and have different lengths, we define a distance to incorporate motion similarity and spatial closeness. For two trajectories A and B with points $\mathbf{x_A}[t]$ and $\mathbf{x_B}[t]$, we calculate their similarity on a temporal overlap $t \in [\tau_1, \tau_2]$ as[1]

$$d(A,B) = \left(\max_{t \in [\tau_1,\tau_2]} \|\mathbf{x_A}[t] - \mathbf{x_B}[t]\|_2 \right) \times \left(\frac{\sum_{t=\tau_1}^{\tau_2} \|\dot{\mathbf{x}}_\mathbf{A}[t] - \dot{\mathbf{x}}_\mathbf{B}[t]\|_2}{(\tau_2 - \tau_1)\sigma_{[\tau_1,\tau_2]}} \right), \qquad (1)$$

where $\dot{\mathbf{x}}[t] = \mathbf{x}[t] - \mathbf{x}[t-1]$ denote velocities of the trajectory points and $\sigma_{[\tau_1,\tau_2]}$ is the local optical flow variance in the interval $[\tau_1, \tau_2]$. In (1), the first term is a measure of spatial distance while the second estimates distance in motion and velocity. To group trajectories, we follow [7] and calculate the affinities between all pairs of trajectories in a video, forming an affinity matrix, calculated as $\omega(A,B) = \exp(-\eta d(A,B))$. A normalized-cut clustering is then used to group the trajectories, where a Cattell's scree test is used to determine the appropriate number of clusters.

Each trajectory group forms a part that may or may not be associated to the activities of interest. For instance, 23 parts appear in the video frame shown in Fig. 3. Each part

[1] Bold capital letters denote matrices (*e.g.*, $\mathbf{D}$). All non-bold letters denote scalar variables. d_{ij} denotes the scalar in the row i and column j of $\mathbf{D}$. $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ denotes the inner product between two vectors $\mathbf{d}_1$ and $\mathbf{d}_2$. $||\mathbf{d}||_2^2 = \langle \mathbf{d}, \mathbf{d} \rangle = \sum_i d_i^2$ denotes the squared Euclidean Norm of $\mathbf{d}$. $||\mathbf{A}||_*$ designates the nuclear norm (sum of singular values) of $\mathbf{A}$.

Fig. 3: Left to right: Points tracked on a frame, extracted trajectories, trajectory groups.

is represented by a histogram of oriented gradients (HoG), optical flow (HoF) [1] and oriented edges in the motion boundaries (HoMB) [5]. These histograms are computed on a regular grid at three different scales. Each descriptor (HoG, HoF, HoMB) uses an independent dictionary, obtained by performing K-means on all the parts, and quantizing all descriptors to its closest ℓ_2 distance dictionary element. The concatenation of all three histograms forms the group (part) descriptor, $\mathbf{h}_k \in \mathbb{R}^n$. A video $\mathbf{V}_i$ is described by concatenating its activity parts, as $\mathbf{V}_i = [\mathbf{h}_{1i}\ \mathbf{h}_{2i}\ \ldots\ \mathbf{h}_{ki}]$.

4 Activity Recognition and Localization

In this section, we present our weakly-supervised learning algorithm for action recognition and localization in video sequences. In our problem, we have several training videos, each of which is labeled with one or more activities. However, no spatio-temporal information exists on where the activities occur. Our task is to classify whether unknown test videos contain those activities or not, and simultaneously localize them throughout the video. Our approach merges the advantages of two recently proposed low-rank models: subspace segmentation [14] clusters similar activity parts from all videos in the dataset, and a matrix completion classifier [38] determines the activity labels they belong to, such that the labeling is consistent throughout the entire dataset.

Let m be the number of different activity classes, n the dimensionality of the feature space, and N_{tr}, N_{tst} the number of training and testing parts, respectively. For the classification task, we can define a matrix $\mathbf{D_0}$ as

$$\mathbf{D_0} = \begin{bmatrix} \mathbf{D_Y} \\ \mathbf{D_X} \\ \mathbf{D_1} \end{bmatrix} = \begin{bmatrix} \mathbf{Y_{tr}} & \mathbf{Y_{tst}} \\ \mathbf{X_{tr}} & \mathbf{X_{tst}} \\ \mathbf{1}^\top & \end{bmatrix}, \tag{2}$$

where $\mathbf{Y_{tr}} \in \mathbb{R}^{m \times N_{tr}}$ and $\mathbf{Y_{tst}} \in \mathbb{R}^{m \times N_{tst}}$ are the training and test labels and $\mathbf{X_{tr}} \in \mathbb{R}^{n \times N_{tr}}$ and $\mathbf{X_{tst}} \in \mathbb{R}^{n \times N_{tst}}$ are the training and test feature vectors, respectively. Hence, $\mathbf{D_Y}$, $\mathbf{D_X}$ and $\mathbf{D_1}$ denote the label, feature and last rows of $\mathbf{D}$, respectively. As noted by Cabral *et al.* [38], if a linear classification model holds, $\mathbf{D_0}$ is rank deficient. Therefore, classification can be posed as a matrix completion problem of filling the missing entries in $\mathbf{Y}_{tst}$ such that the nuclear norm of $\mathbf{D_0}$ (a convex approximation of its rank) is minimized. To deal with noise and outliers in the data, we can incorporate an error term $\mathbf{E^{mc}}$ in the known feature and training label entries,

$$\mathbf{D} = \mathbf{D_0} + \mathbf{E^{mc}} = \begin{bmatrix} \mathbf{Y_{tr}} & \mathbf{Y_{tst}} \\ \mathbf{X_{tr}} & \mathbf{X_{tst}} \\ \mathbf{1}^\top & \end{bmatrix} + \begin{bmatrix} \mathbf{E_{Y_{tr}}} & \mathbf{0} \\ \mathbf{E_X} & \\ \mathbf{0}^\top & \end{bmatrix} \tag{3}$$

and the classification process can be posed as finding the best $\mathbf{Y_{tst}}$ and the error matrix $\mathbf{E^{mc}}$ such that the rank of $\mathbf{D}$ is minimized.

As discussed in Section 3, each video $\mathbf{V}_i$ is represented by the histograms of its activity parts. If labels were provided for each part in training, we could construct $\mathbf{D_0}$ by setting each column to the features corresponding to one activity part and its respective $\{0,1\}^m$ label vector. However, in our case supervision is weak and labels are only provided for entire videos. Thus, simply labeling parts with all class labels present in the video they originate from is insufficient for obtaining correct part level classifications.

Instead, to identify the parts that comprise each activity class, we can also exploit the fact that activity parts from the same class likely cluster together. This can be formulated as a segmentation of feature vectors into low-rank subspaces, using a Low-Rank Representation (LRR) [14]. Since $\mathbf{D_X}$ contains the feature vectors for all videos in the dataset, we can cluster activity parts by computing a low-rank similarity matrix $\mathbf{Z}$, as

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{E^{lrr}}} \quad & \|\mathbf{Z}\|_* + \lambda \|\mathbf{E^{lrr}}\|_{2,1}, \\ \text{subject to} \quad & \mathbf{D_X} = \mathbf{D_X Z} + \mathbf{E^{lrr}}, \end{aligned} \tag{4}$$

where $\mathbf{E^{lrr}}$ is the LRR [14] error matrix and λ is a balancing parameter between low-rank and error fit. $\mathbf{Z}$ is indicative of the similarity between each activity part in $\mathbf{D_X}$ and thus can be used as an additional cue to weak supervision for classifying which parts constitute which activities. Using the similarity matrix $\mathbf{Z}$, we can apply a clustering method such as Normalized Cuts to group similar activity parts in all train/test videos. The output of this clustering method is a $n_c \times N$ binary matrix $\mathbf{Q}$, where n_c is the number of clusters. Each row of $\mathbf{Q}$ corresponds to one cluster, with $q_{ij} = 1$ if the j^{th} activity part belongs to the i^{th}cluster, and 0 otherwise.

Below, we show that these matrix completion classification and subspace clustering steps can be done jointly, so that labels are consistent within clusters and vice-versa.

4.1 Joint Classification and Clustering

With the matrix completion and subspace segmentation defined as above, we can simultaneously obtain a low-rank representation of the feature vector matrix $\mathbf{D_X}$, and correct and complete the labels in $\mathbf{D_Y} = [\mathbf{Y_{tr}}, \mathbf{Y_{tst}}]$. Our activity classification problem can be defined as minimizing the rank of $\mathbf{D}$ for determining the part labels, while at the same time ensuring the labels are consistent with the clustering $\mathbf{Q}$ obtained from the low-rank representation $\mathbf{Z}$ of the parts $\mathbf{D_X}$. If we define $\mathbf{\Omega_Y}$ as the set of known label entries in $\mathbf{D_0}$, this objective can be written as

$$\begin{aligned} \min \quad & \|\mathbf{D}\|_* + \gamma \|\mathbf{Z}\|_* + \lambda \|\mathbf{E^{lrr}}\|_{2,1} \\ & + \rho_1 \sum_{i,j \in \mathbf{D_Y}} c_y(d_{ij}, q_{kj}) + \rho_2 \sum_{i,j \in \mathbf{\Omega_Y}} c_y(d_{ij}, d_{0ij}) \\ \text{subject to} \quad & \mathbf{D} = \mathbf{D_0} + \mathbf{E^{mc}}, \mathbf{D_1} = \mathbf{1}^\top, \mathbf{D_X} = \mathbf{D_X Z} + \mathbf{E_X}, \end{aligned} \tag{5}$$

where $c_y(a,b) = \log\left(1 + \exp\left(-(2b-1)(a-b)\right)\right)$ is a logistic loss function that penalizes entries of different classes. γ, λ, ρ_1, ρ_2 are positive trade-off parameters. k is the most similar cluster to label i, calculated as $k = \operatorname{argmin}_{k=1}^{nc} \sum_j c_y(d_{ij}, q_{kj})$.

With the objective in (5), the first term seeks a low-rank $\mathbf{D}$ matrix so that labels can be expressed as a linear combination of features. The second establishes a low-rank representation $\mathbf{Z}$ for subspace clustering. The third term controls the level of noise in the clustering. The fourth term nudges the labels in $\mathbf{D_Y}$ the direction suggested by the clustering $\mathbf{Q}$ and the fifth term regularizes changes on known training labels $\mathbf{Y_{tr}}$ in the matrix completion. Therefore, we are seeking to achieve a consensus between the clustering and classification outputs. The intersection of these two tasks is incorporated by the fourth term, where inconsistent clustering outputs and labels are penalized. The minimization process will aim towards unanimity between the two and the least label changing in $\mathbf{Y_{tr}}$. Also, notice that in the process of joint minimization, both classification and clustering tasks share the feature error matrix, resulting in less variables than used when optimizing both objectives separately.

The objective in (5) can be optimized using an Alternating Direction Method of multipliers (ADMM) [39]. When it converges, the labels in $\mathbf{Y_{tst}}$ corresponding to each activity part indicate its action label(s) and the columns with that label are the parts associated to that specific activity. The highest computational complexity step in solving (5) with an ADMM is a SVD of $\mathbf{D}$, but scalable SVD/ADMM methods are currently being researched heavily [40].

As in $\mathbf{D_Y}$, each instance is assigned a set of labels, each of which belongs to an independent activity class. This enables us to model multi-label MIL problems. Many previous works have exploring the dependence among the labels [41, 42]. But when the labels are incomplete (weakly-supervised) the task is harder. As also explored in previous works [18, 38], the low rank assumption of the matrix $\mathbf{D}$ resembles a linear dependence among the labels and the feature vectors. We evaluate our multi-label setting in a weakly-supervised video activity recognition and localization.

5 Experiments

To evaluate the proposed technique, we set up several experiments on various synthetic and real datasets. Since our approach performs clustering and classification simultaneously, one might conceive that we could first run clustering and then use matrix completion for obtaining the labels. Thus, as a baseline, we derive a low-rank representation [14] of matrix $\mathbf{D_X}$ and then run matrix completion while incorporating the feature error term in the matrix completion formulation (LRRMC).We also compare the performance of our method to using just matrix completion (MC) of [38] for classification as described in Sec.4 to show that solely relying on a weakly supervised labeling for part classification does not work, and the well-known MI-SVM [33], with RBF kernel.

In each iteration of (5), we obtain the clustering $\mathbf{Q}$ using $n_c = 2m$ clusters to account for intra-class variability, and use as parameters $\gamma = 0.9$, $\rho_1 = 1.5$, $\rho_2 \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$. For experiments on activity recognition datasets, to ensure direct comparability with state of the art methods, we follow the setup of [7] for obtaining and describing activity parts, as described in Sec. 3. Each part is represented by histogram of oriented gradients (HoG), histogram of optical flow (HoF) [1] and histogram of the oriented edges in the motion boundaries (HoMB) [5] descriptors, with $500, 500, 300$ dimensions respectively.

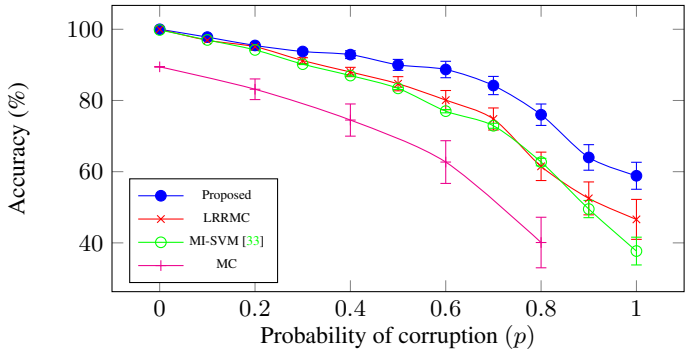


Fig. 4: Accuracy comparison according to corruption probability p on synthetic data. This figure shows the means and standard deviations for three different runs.

5.1 Synthetic Data

First, in order to validate the proposed algorithm, we construct 10 independent subspaces of dimensionality 100 (as described in [14]). The first five subspaces form our desired positive classes and the second five, negative. We create 100 positive and 100 negative bags, with size 10, and sample instances from the above subspaces. Positive bags, as in MIL, are composed of uniformly distributed positive and negative instances. We corrupt each sampled instance $\mathbf{x}$ with probability p, by adding Gaussian noise with zero mean and variance $0.3\|\mathbf{x}\|$. The performance of the proposed method is compared with LRRMC, MI-SVM and matrix completion (MC) [38], as illustrated in Fig. 4 for different probabilities of corruption and noise. The performance of our method is much better when the noise level increases in the data. As mentioned in Sec. 4, MC yields worse results since it fully relies on the initial labeling, which is not accurate enough due to its weakly supervised nature. Our method performs a joint clustering and classification of the data and detects noise and outliers in both tasks collaboratively. In LRRMC these are done separately. Thus, our method deals better with noise in the data.

5.2 Action recognition and localization

Three popular activity recognition datasets are used: MSR-II [6], HOHA [1] and UCF sports [3] action datasets. MSR-II action dataset 2 contains 54 videos with three action categories: boxing, clapping and hand-waving. In this dataset, some of the videos contain multiple actions and some with actions even occurring at the same time. The HOHA (Hollywood1 Human Action) dataset contains 430 videos. Each video contains significant camera motion, rapid scene changes and occasionally significant clutter. Furthermore, actions in this dataset are performed in different conditions, and many actions are defined by the interactions between the subjects and/or objects. These factors make this dataset particularly challenging. The UCF sports dataset consists of 150 videos extracted from sports broadcasts. Video in this dataset contain camera motions and many different lighting and capturing conditions, as well as large displacements of most of the actions, cluttered backgrounds, and large intra-class variability.

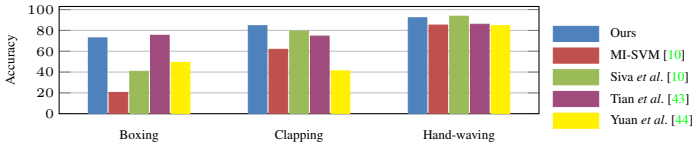


Fig. 5: Per-class recognition accuracy for MSR-II dataset.

Table 1: Recognition results on MSR-II dataset. *Cross dataset* methods are trained on KTH dataset, which only contains actions with little background motion.

Method	**Supervision**	**Accuracy**
Siva *et al*. [10]	Weak	71.2%
MI-SVM [10]	Weak	55.8%
Tian *et al*. [43]	Full (Cross dataset)	78.8%
MC	Weak	41.1%
LRRMC	Weak	54.9%
Our Method	Weak	**83.1**%

Recognition: Tests on each of the datasets have separate experimental settings to facilitate comparisons with reference methods. We compare our recognition model with state-of-the-art models reported in the literature and with the same baselines described in the synthetic tests of Sec. 5.1. The final classification step in our model is performed via a thresholding procedure, where labels above a common threshold are selected.

MSR-II dataset - For the experiments on this dataset, a two-to-one random division of all videos in the dataset creates the training and testing sets. This dataset contains videos with multiple actions happening in the video and, in some cases, being performed at the same time, which can challenge our multi-label classification framework. Some of the videos in this dataset contain several instances of all activities. Since we expect a single instance of each activity class in the video, the videos are split such that each video contains only one instance of each activity class, but allowing for several activities from different classes. Fig. 5 shows our per-class accuracy results compared to the MI-SVM model [33]. Table 1 shows the recognition accuracy results compared to state-of-the-art methods on this dataset. The supervision column shows the level of supervision used in the training phase: fully supervised methods know spatio-temporal bounding boxes of activity locations, whereas weakly-supervised methods use only the label(s).

HOHA dataset - In this experiment the test set has 211 videos with 217 labels and the training set has 219 videos with 231 labels, all manually annotated [7]. Fig. 6 shows the per-class accuracy results for this dataset. This dataset is very challenging for activity recognition, due to the large amount of clutter and motion in the camera. Our approach is comparable with results from state-of-the-art methods designed specifically for this dataset, improving them by a slight margin. Table 2 gives the overall accuracy results compared to some other methods on this dataset.

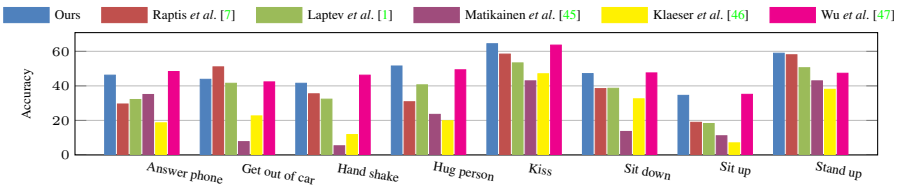


Fig. 6: Per-class recognition for HOHA dataset.

Table 2: Recognition results on HOHA dataset.

Method	Supervision	Accuracy
Klaeser *et al.* [46]	Full	27.3%
Laptev *et al.* [1]	Full	38.4%
Matikainen *et al.* [45]	Full	22.8%
Raptis *et al.* [7]	Full	40.1%
Wu *et al.* [47]	Full	47.6%
MC	Weak	22.3%
LRRMC	Weak	29.8%
Our Method	Weak	**48.5**%

UCF Sports dataset - We split this dataset into 103 training and 47 test samples, follwing the setup described in [7, 8]. This separation minimizes the strong correlation of background cues between the testing and training set [7]. Some results on this dataset report leave-one-out-cross-validation (LOOCV) performance, which may take into account the similarity of the background instead of the activity itself. In this dataset the background is very similar for sports of the same kind, which affects the activity recognition rates. Fig. 7 depicts the per-class classification accuracy for this dataset. As shown, our method outperforms the BoW+SVM model in almost all classes. As shown in Table 3, the overall recognition rate of our method is also competitive with the state-of-the-art. The upper part of the table compares our results with state-of-the-art methods' reported results for the same training and testing dataset split. Our method outperforms all of these works. The lower part of the table shows results from works that use LOOCV, which generally achieve better results. Our split is much harder and the difference between the results is expected. Notwithstanding a more difficult test scenario, our results are still comparable to these works.

Spatio-temporal localization: The second function of our method is the spatio-temporal localization of the activity in the video sequence. In order to assess spatio-temporal localization directly against reported state-of-the-art methods, we employ three metrics for assessing localization performance: 1) intersection-over-union using the selected positive parts (IOU), 2) average precision (AP) of part classification based on ground truth spatio-temporal annotations, and 3) the localization $score$, defined as in [7]. The latter is defined as the average ratio of the sets of points inside the annotated ground truth bounding box and the set of points of the selected trajectory group for each frame. If the detected activity part(s) throughout the video have at least a θ overlap with

Fig. 7: Per-class recognition results for UCF Sports dataset.

Table 3: Recognition results on UCF Sports. Upper part: Results with 103:47 dataset split. Lower part: Results with LOOCV.

Method	Supervision	Accuracy
Lan *et al.* [8]	Full	73.1%
Raptis *et al.* [7]	Full	79.4%
Tian *et al.* [12]	Full	75.2%
Ma *et al.* [31]	Weak	81.7%
MC	Weak	59.8%
LRRMC	Weak	71.2%
Our Method	Weak	**86.9**%
Le *et al.* [48]	Full	86.5%
Wang *et al.* [20]	Full	85.6%
Wang *et al.* [5]	Full	88.2%
Wang *et al.* [49]	Full	**89.1**%
Kovashka and Grauman [21]	Full	87.3%

Table 4: Action localization AP on the MSR-II dataset. *Cross dataset* methods are trained on KTH dataset, which only contains actions with little background motion.

Method	Supervision	Clapping	Boxing	Handwaving
Siva *et al.* [10]	Full	**0.602**	0.694	0.700
Siva *et al.* [10]	Weak	0.326	0.658	0.799
Cao *et al.* [6]	Full (Cross Dataset)	0.125	0.144	0.242
Tian *et al.* [12]	Full (Cross Dataset)	0.239	0.389	0.447
Our Method	Weak	0.569	**0.724**	**0.811**

the annotated ground truth bounding box ($score \geq \theta$), the recognition/localization is considered as correct. The results are compared to the state-of-the-art methods in the literature, using IOU, AP or localization score, where available. Tables 4, 5 and 6 show results on MSR-II, HOHA and UCF Sports datasets, respectively. Since [8] only provides localization results on a subset of frames, we also include results on this subset for comparison. The average recognition/localization accuracies for the experiments on the datasets as a function of θ are illustrated in Fig. 8. Some results are shown in Fig. 9.

Experimental results discussion: Our experiments show that the proposed joint process in (5) significantly improves results, when compared to the baselines of MC and performing clustering and classification steps separately (LRRMC). We note that

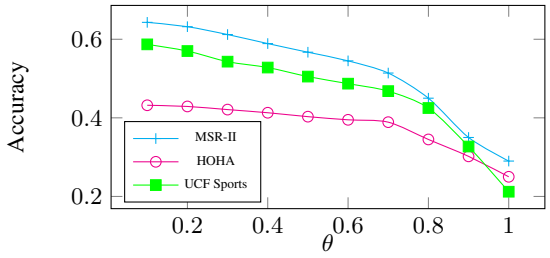


Fig. 8: Average localization accuracy as a function of the localization overlap θ.

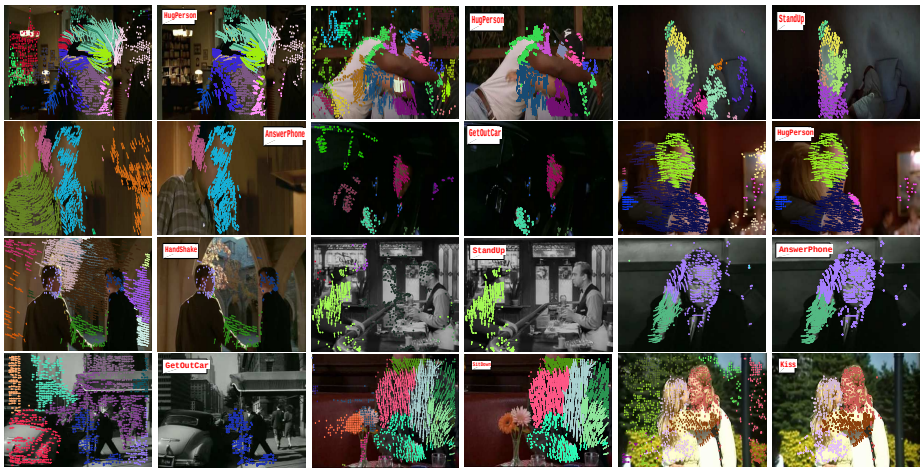
Table 5: Localization comparisons for HOHA dataset.

Method	**Supervision**	**Localization score**		**Mean IOU**
		$\theta = 0.1$	$\theta = 1$	
Raptis *et al.* [7]	Full	54.3%	**28.6%**	–
Our Method	Weak	**56.2%**	21.0%	**42.9%**

Table 6: Average localization IOU on the UCF Sports dataset. Note that [26] and [8] use the bounding box annotations during the training, while ours is weakly-supervised.

Action	**Subset of frames**				**All frames**			
	[26]	[8]	[31]	Ours	[26]	[8]	[31]	Ours
Diving	36.5	43.4	**46.7**	44.8	37.0	–	**44.3**	43.7
Golf	–	37.1	51.3	**53.1**	–	–	50.5	**52.3**
Kicking	–	36.8	50.6	**54.3**	–	–	48.3	**52.9**
Lifting	–	68.8	55.0	**69.0**	–	–	51.4	**63.5**
H-Ride	**68.1**	21.9	29.5	34.5	**64.0**	–	30.6	32.5
Running	**61.4**	20.1	34.3	31.2	**61.9**	–	33.1	30.1
Skating	–	13.0	40.0	**45.5**	–	–	38.5	**43.2**
Swing-B	–	32.7	54.8	**57.1**	–	–	54.3	**57.5**
Swing-S	–	16.4	19.3	**48.7**	–	–	20.6	**44.1**
Walking	–	28.3	39.5	**47.5**	–	–	39.0	**47.1**
Avg.	–	31.8	42.1	**51.3**	–	–	41.0	**46.7**

the multi-label nature of our method allows us to provide results for simultaneous actions on the MSR-II dataset, as seen on Fig. 9. An important note on the recognition results, is that our method performed competitively even with those specifically focused for recognition (*i.e.*, that do not perform any localization of the activity) and methods that train with fully annotated datasets. This is despite the fact that when using the whole frame or video features for recognition, we are dealing with many outliers and significant noise. Furthermore, our model extracts the exact spatio-temporal segmentation of the activity, rather than a simple bounding box, cuboid or voxel representation, as opposed to many previous works. We improve the recognition results on all datasets, and also achieve good localization scores. We believe these could be improved further



(a) HOHA Dataset

(b) UCF Sports Dataset

(c) MSR-II Action Dataset

Fig. 9: Recognition and localization results on action recognition datasets. Each result from a test video is illustrated in a pair of images, first of which is a sample frame of the video containing the action of the interest. The trajectory groups are shown on this image, each with a different color. The second image shows the selected trajectory group(s) by our algorithm. (a) results from the HOHA dataset, (b) results from the UCF Sports dataset, and (c) results from MSR-II action datset.

if more accurate spatio-temporal annotations in the datasets were used as ground truth instead of bounding boxes.

As could be seen, our method achieved much better results compared to many state-of-the-art methods. This is basically due to two important properties of our method. Our method deals with errors and outliers in the feature vectors and the labels. As could be seen in (5) we extract the erroneous elements as well in the process of minimizing the matrix ranks. The error for both LRR and MC are incorporated simultaneously, which tend to correct one another in the process. On the other hand, our method labels the actions via transduction, which alone improves the results compared to inductive approaches. There are no separate train and test phases and our approach incorporates activity parts and information from the whole dataset when minimizing the nuclear norm and deciding on the instance classes.

6 Conclusions

In this paper, we have proposed a low-rank formulation for weakly supervised learning and have applied it to the challenging problem of activity recognition. Our approach uses a simultaneous convex matrix completion and LRR subspace clustering framework to recover the labels for the test videos and localize the spatio-temporal extent of activities throughout each video. Interactions between the activity parts are globally modeled throughout the entire dataset using the subspace clustering procedure, while the matrix completion framework labels the activities ensuring that labeling is consistent within clusters and vice-versa. Our experiments show this joint process significantly improves results, when compared to performing clustering and classification steps separately. Moreover, it attains performances comparable to state-of-the-art methods for classification and localization in all three datasets tested.

Unlike typical MIL approaches, our method to be naturally multi-label and is able to handle video sequences where several activity parts have to occur together in a bag to define an action, and actions occur simultaneously in different spatial locations.

As a direction for future work, we intend to apply and develop incremental procedures for the training and testing and exploit parallel algorithms for the SVD operations needed to optimize (5), such as in [40], in order to decrease processing time.

References

[1] Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008)
[2] Ryoo, M.S., Aggarwal, J.K.: Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV. (2009)
[3] Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR. (2008)
[4] Sheikh, Y., Sheikh, M., Shah, M.: Exploring the space of a human action. In: ICCV. (2005)
[5] Wang, H., Kläser, A., Schmid, C., Cheng-Lin, L.: Action Recognition by Dense Trajectories. In: CVPR. (2011)

[6] Cao, L., Liu, Z., Huang, T.S.: Cross-dataset action detection. In: CVPR. (2010)
[7] Raptis, M., Kokkinos, I., Soatto, S.: Discovering discriminative action parts from mid-level video representations. In: CVPR. (2012)
[8] Lan, T., Wang, Y., Mori, G.: Discriminative figure-centric models for joint action localization and recognition. In: ICCV. (2011)
[9] Nguyen, M.H., Torresani, L., De la Torre, F., Rother, C.: Weakly-supervised discriminative localization and classification: a joint learning process. In: ICCV. (2009)
[10] Siva, P., Xiang, T.: Weakly-supervised action detection. In: BMVC. (2011)
[11] Zhou, Z., Zhang, M.: Multi-instance multi-label learning with application to scene classification. In: NIPS. (2006)
[12] Tian, Y., Sukthankar, R., Shah, M.: Spatiotemporal deformable part models for action detection. In: CVPR. (2013)
[13] Wang, L., Qiao, Y., Tang, X.: Motionlets: Mid-level 3d parts for human motion recognition. CVPR (2013) 2674–2681
[14] Liu, G., Lin, Z., Yu, Y.: Robust subspace segmentation by low-rank representation. In: ICML. (2010)
[15] Cheng, B., Liu, G., Wang, J., Huang, Z., Yan, S.: Multi-task low-rank affinity pursuit for image segmentation. In: ICCV. (2011)
[16] Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: CVPR. (2009)
[17] Tang, K., Sukthankar, R., Yagnik, J., Fei-Fei, L.: Discriminative segment annotation in weakly labeled video. In: CVPR. (2013)
[18] Goldberg, A.B., Zhu, X., Recht, B., Xu, J.M., Nowak, R.D.: Transduction with matrix completion: Three birds with one stone. In: NIPS. (2010)
[19] Feng, S., Xu, D.: Transductive multi-instance multi-label learning algorithm with application to automatic image annotation. Expert Syst. Appl. **37** (2010) 661–670
[20] Wang, H., Ullah, M.M., Kläser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC. (2009)
[21] Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: CVPR. (2010)
[22] Hoai, M., Lan, Z., De la Torre, F.: Joint segmentation and classification of human actions in video. In: CVPR. (2011)
[23] Shapovalova, N., Vahdat, A., Cannons, K., Lan, T., Mori, G.: Similarity constrained latent support vector machine: An application to weakly supervised action classification. In: ECCV. (2012) 55–68
[24] Tang, K., Fei-Fei, L., Koller, D.: Learning latent temporal structure for complex event detection. In: CVPR. (2012)
[25] Chen, C.Y., Grauman, K.: Efficient Activity Detection with Max-Subgraph Search. In: CVPR. (2012)
[26] Tran, D., Yuan, J.: Max-margin structured output regression for spatio-temporal action localization. In: NIPS. (2012)
[27] Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: ICCV. (2009)
[28] Tran, D., Yuan, J., Forsyth, D.: Video event detection: From subvolume localization to spatio-temporal path search. IEEE Trans. PAMI (2013)

[29] Kumar, B.G.V., Patras, I.: Supervised dictionary learning for action localization. In: FG. (2013)
[30] Gaidon, A., Harchaoui, Z., Schmid, C.: Temporal localization of actions with actoms. IEEE Trans. PAMI **35** (2013) 2782–2795
[31] Ma, S., Zhang, J., Ikizler-Cinbis, N., Sclaroff, S.: Action recognition and localization by hierarchical space-time segments. In: ICCV. (2013)
[32] Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. AI **89** (1997) 31–71
[33] Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS. (2003)
[34] Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: CVPR. (2012)
[35] Hartmann, G., Grundmann, M., Hoffman, J., Tsai, D., Kwatra, V., Madani, O., Vijayanarasimhan, S., Essa, I., Rehg, J., Sukthankar, R.: Weakly-supervised learning of object segmentations from web-scale video. In: ECCV. (2012)
[36] Li, F., Sminchisescu, C.: Convex multiple-instance learning by estimating likelihood ratio. In: NIPS. (2010)
[37] Joulin, A., Bach, F.: A convex relaxation for weakly-supervised classifiers. In: ICML. (2012)
[38] Cabral, R.S., De la Torre, F., Costeira, J.P., Bernardino, A.: Matrix completion for multi-label image classification. In: NIPS. (2011)
[39] Lin, Z., Chen, M., Wu, L., Ma, Y.: The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. UIUC technical report 2215 (2009)
[40] Tron, R., Vidal, R.: Distributed computer vision algorithms through distributed averaging. In: CVPR. (2011)
[41] Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition **37** (2004) 1757 – 1771
[42] Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE T KDE **99** (2013)
[43] Tian, Y., Cao, L., Liu, Z., Zhang, Z.: Hierarchical filtered motion for action recognition in crowded videos. IEEE Trans. Sys., Man, and Cyb., Part C **42** (2012) 313–323
[44] Yuan, J., Liu, Z., Wu, Y.: Discriminative video pattern search for efficient action detection. IEEE Trans. PAMI **33** (2011) 1728–1743
[45] Matikainen, P., Hebert, M., Sukthankar, R.: Trajectons: Action recognition through the motion analysis of tracked features. In: ICCV. (2009)
[46] Klaeser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC. (2008)
[47] Wu, S., Oreifej, O., Shah, M.: Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In: ICCV. (2011)
[48] Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR. (2011)
[49] Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. IJCV (2013)