Similarity Constrained Latent Support Vector Machine: An Application to Weakly Supervised Action Classification [*]

Nataliya Shapovalova, Arash Vahdat, Kevin Cannons,
Tian Lan, and Greg Mori

School of Computing Science, Simon Fraser University, Canada
{nshapova,avahdat,kcannons,tla58,mori}@cs.sfu.ca

Abstract. We present a novel algorithm for weakly supervised action classification in videos. We assume we are given training videos annotated only with action class labels. We learn a model that can classify unseen test videos, as well as localize a region of interest in the video that captures the discriminative essence of the action class. A novel Similarity Constrained Latent Support Vector Machine model is developed to operationalize this goal. This model specifies that videos should be classified correctly, and that the latent regions of interest chosen should be coherent over videos of an action class. The resulting learning problem is challenging, and we show how dual decomposition can be employed to render it tractable. Experimental results demonstrate the efficacy of the method.

1 Introduction

Max-margin parameter learning for latent variable models is a popular approach in object and action recognition. The Latent SVM formalism has been successfully applied for many tasks including object detection, action recognition, and human pose estimation (e.g. [1,2,3,4]). One reason for their success is that they allow for weak supervision of latent parts or subregions within an element to be recognized. For instance, in action recognition from video data, a weakly supervised method might only specify the action class label of a training video, rather than providing manually labeled localization or region of interest data. The latent SVM can model the localization of the action within the video, but does not require its annotation for learning.

[*] This work was supported by a Google Research Award, MITACS-Elevate and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VII, LNCS 7578, pp. 55–68, 2012.

However, the performance of these methods often depends on a good initialization of the latent variables. In this paper we develop a novel learning framework that considers pairwise similarity of the latent variables. Conceptually, enforcing pairwise similarity of the latent variables allows for greater consistency of the latent variables across the training set and is a natural way to define them.

In this paper, we present a novel extension of the popular Latent SVM [1]. This new learning framework is subsequently applied to the task of action classification [5]. Specifically, given an input video, we would like to automatically determine whether a particular action is taking place. Such a binary classification decision is useful for many applications. However, it is often desirable to know more – for example where does the action occur, or what video evidence is leading to the action classification decision. We treat the spatio-temporal location of this video evidence as a latent variable in our model.

Building systems that identify such discriminative evidence typically requires hand-labeling of regions of interest on training videos. But this is an arduous, time-consuming, and error-prone process. Deploying such methods to large datasets would be costly. Furthermore, labeling where an action takes place, or what visual evidence is indicative of a particular action is a non-trivial task. The spatio-temporal boundaries of an action are difficult to define. Context, in the form of a background scene or relevant objects, also often plays a crucial role in action classification, and should arguably be included in a region of interest localization.

In this paper, we propose such a learning algorithm that produces a novel weakly supervised method for action classification. Fig. 1 provides an overview of our approach. Our work is inspired by two recent pieces of work in the vision literature. Lan et al. [2] showed that reasoning about a latent region of interest can improve action recognition results. However, that method requires supervision of the latent regions of interest on the training data. We believe relaxing that assumption is an important direction for deploying such methods more widely. Alexe et al. and Vezhnevets et al. [6,7] have done impressive work on automatic object localization and segmentation from weakly supervised data. An "objectness" saliency operator is used to guide the weakly supervised algorithm to localize objects of a consistent appearance. This line of work presents two-stage processes in which the object and classification or segmentation are not learned together. It is typically the case that improved performance results if these are learned jointly. Here we build on this line of work, and present a unified discriminative framework for jointly learning a classifier and localizing discriminative regions in video.

The main technical contribution of this paper is the development of the *Similarity Constrained Latent SVM*, a formalism for this type of problem. At a high level, this formalism allows for latent variables (action evidence locations or regions of interest) similar to the popular latent support vector machine [1]. However, it adds the ability to encourage consistency of the latent variables across all of the training data, considering pairwise similarities of latent labels

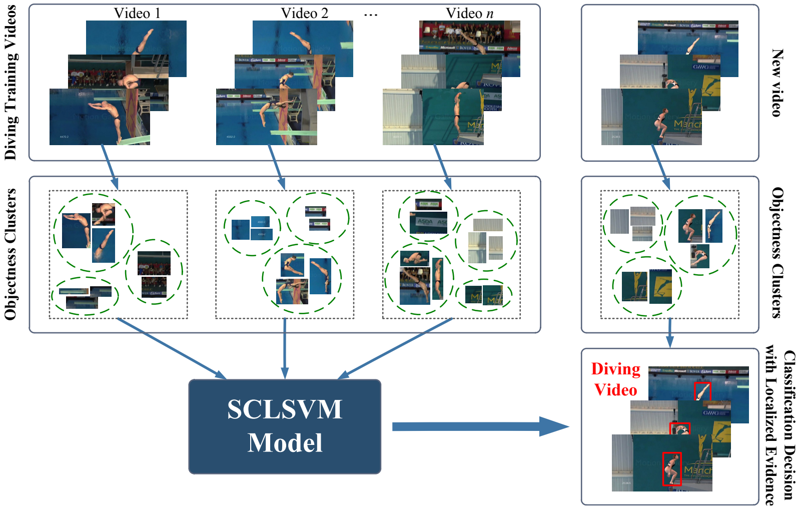


Fig. 1. Overview of our approach. First, for each video, we extract ROI candidates by applying the objectness operator and mean shift clustering. Second, we employ the Similarity Constrained Latent SVM learning framework to produce a corresponding action classifier. Finally, we apply the learnt classifier to the test videos; as output, we get the action label of the video as well as a specific ROI, which serves as the evidence of the action.

in a fashion similar to the Transductive SVM [8] for semi-supervised learning[1]. A particular technical challenge in this formulation is that a desire for consistency of latent regions across all training data leads to a more complex learning procedure. We show that approximate inference based on dual decomposition [9,10] can be used to address this issue.

2 Previous Work

In this paper we develop a novel learning algorithm that is applied to the problem of weakly supervised action classification with evidence localization. The aforementioned work on objectness [6,7] and joint localization and recognition [2] are most closely related to the work presented here. More broadly, weakly supervised methods have been explored in the vision literature, with a particular focus on object recognition.

Fergus et al. [11] develop weakly supervised methods for object class recognition. A probabilistic part-based model is used, with supervision only at the object class level rather than parts. The latent SVM object detector of Felzenszwalb et al. [1] is similar – class level supervision is provided (e.g. an image window containing a person), without part locations (e.g. the positions of body parts). Viola et al. [12] developed a similar method for face detection, capturing

[1] This model will be applied inductively to test data, unlike the Transductive SVM.

the variability in ground-truth labeling of face locations in training data, using a boosting framework. Yao and Fei-Fei [3] and Yang et al. [13] perform action recognition in images. Weak supervision of human body pose and objects are considered in these frameworks. Laptev et al. [14] build datasets for action recognition by considering surrogate movie script data, a form of weak supervision about action locations in videos. Bilen et al. [4] perform latent localization of objects and actions in a max-margin framework.

Alternatives such as crowd-sourcing (e.g. [15]) could be used to provide detailed labeling, or certainly to aid in narrowing a search for appropriate latent regions. However, the problems mentioned above, of accurate spatio-temporal labeling of regions of interest, and decisions about contextual regions still persist. The efficacy of the weakly supervised methods for object and action recognition suggest this direction is fruitful.

Our work formulates a max-margin latent variable model [1,16,17] for action recognition with evidence localization, a common paradigm in the vision literature. We build on this formalism to include a loss function that ties the latent variables over all the training videos to have a similar appearance. This approach is related to the max-margin clustering of Xu et al. [18] and the Transductive SVM for structured variables of Zien et al. [19]. The formulation we present follows in the line of Zien et al., yet uses latent variables rather than the unobserved class labels in their semi-supervised approach. The resulting learning problem presents a challenging inference task of jointly inferring latent regions for all training images at once. We use dual decomposition [9] to approximately solve this inference. Dual decomposition has previously been applied to Markov Random Field parameter learning in vision by Komodakis [10].

3 Latent Region of Interest Model

The goal in this paper is to develop a novel learning framework for application to action classification in video sequences. This model should produce accurate classification of new videos, and in addition produce a region of interest within each video that captures the discriminative essence of the action class. In this section we describe the form of this model, the features we use, and the representation for the latent region of interest.

3.1 Video Representation

We represent videos using a local feature approach. Statistical representations of the occurrence of local visual features have proven effective for action recognition (e.g. [14]). Beyond the standard bag-of-words approach to characterizing a video, we will also include a latent region of interest within a video sequence. The final representation for a video is the concatenation of these two representations. For each action class, we will learn a set of parameters that describe which local features tend to appear in entire videos and within latent regions of interest of videos belonging to that class.

More formally, each video x is to be classified with an action label y. We will formulate a model, with parameters w, for scoring a video x with a class label y:

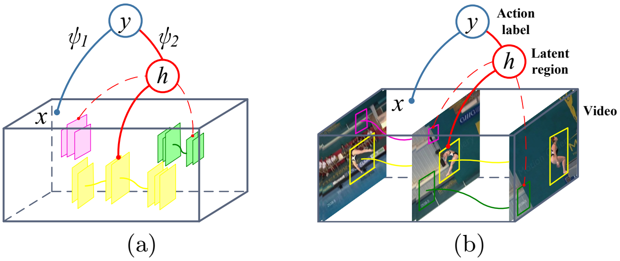
$$F_w(x,y) = \max_h f_w(x,y,h) \tag{1}$$

$$f_w(x,y,h) = w^T \Psi(x,y,h) = w_1^T \psi_1(x,y) + w_2^T \psi_2(x,h,y) \tag{2}$$

In this form, $\psi_1(x,y)$ is a feature vector extracted from the whole video x. Here we use a statistical bag-of-words style representation, computing a bag-of-words histogram of local features. We use a standard approach, with densely sampled HOG3D descriptors [20] vector quantized using k-means. The notation $\psi_1(x,y)$ allows for different components of w_1 to be active for different class labels, a linear model for each class y.

The second model component $\psi_2(x,h,y)$ is a similar feature, but limited in scope to a latent region of interest specified by h. The latent region of interest specifies a spatio-temporal sub-region of a video. The potential function $\psi_2(x,h,y)$ aggregates a bag-of-words histogram on the same HOG3D features, but only over the latent sub-region of the video.

An illustration of this model is presented in Fig. 2. In the next section we describe the particular choice of latent regions we use in our experiments. However, it should be emphasized that the learning framework we develop is general-purpose, and could be used with other feature representations and latent variable representations.



(a) (b)

Fig. 2. A graphical illustration (a) and an example (b) of our model with latent regions. A latent variable h selects a subregion of a video x. Descriptors are computed over the entire video (ψ_1) and the selected subregion (ψ_2), and used to predict action label y.

3.2 Generating Candidate ROIs

Given an input video, there is a huge combinatorial set of potential spatio-temporal regions of interest that one could consider. The learning and inference algorithms we describe will need to operate over this set. Hence, devising a strategy to limit or optimize the search for latent regions is of primary importance.

Our approach is to build a reduced set of candidate regions of interest in the video. We utilize the "objectness" operator developed by Alexe et al. [6]. This operator works on a single frame of the video, and acts like an interest operator, returning bounding boxes that are likely to contain objects. Since the variety of objects and human poses one might see in videos is highly variable, we use a very low threshold on the objectness operator to return a large set of possible bounding boxes. Examples of these are shown in Fig. 3.

We would like to consider "tracklets" of these interesting regions over time to construct candidate regions of interest. However, one cannot rely on straightforward tracking since the region might change appearance dramatically, might appear and disappear over the course of the video, or be poorly localized via the objectness operator.

Instead, we perform an appearance-based clustering of all the objectness bounding boxes. The intuition is that groups of similar bounding boxes will appear in clusters and form a hypothesis about the ROI for the video.

We use mean shift clustering on RGB color histograms representing each objectness bounding box. This results in tens of clusters for each video. Fig. 3 shows examples of these clusters. For our method to be successful, it is necessary that a reasonable ROI must be present within this set of candidates. In our experiments, this procedure was effective, and nearly always generates at least one qualitatively "good" candidate region of interest in each video.

We will then represent a video using a global bag-of-words histogram, combined with a bag-of-words histogram focused on one of these potential regions of interest. Note that the choice of regions of interest based on objectness and mean-shift clustering is a specific choice to ground the description. The method we describe is general, and can be applied to a variety of descriptors and latent ROI representations, ranging from clips of frames to segmentations. We next describe how to learn the model parameters for this representation.

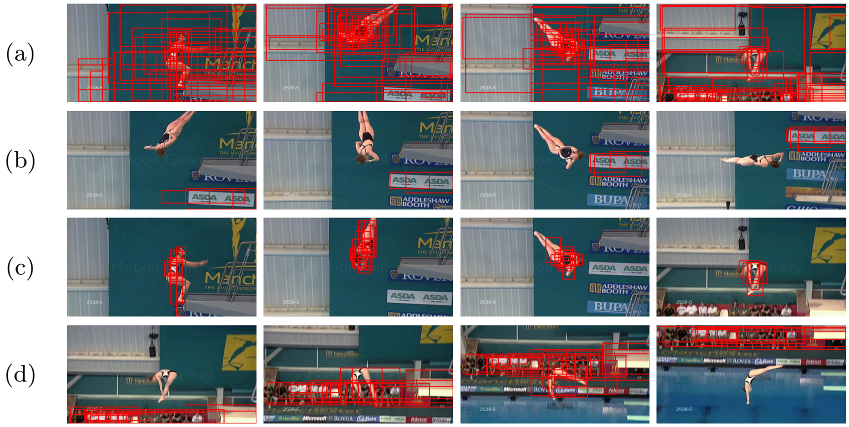


Fig. 3. An illustration of extracting candidate regions of interest. (a) Examples of objectness on frames of a video. (b)-(d) three clusters of objectness bounding boxes grouped by clustering with mean shift. Typically, each cluster captures a tracklet of an object or a group of objects in time.

4 Similarity Constrained Latent SVM

We now define the *Similarity Constrained Latent SVM*, a learning algorithm for weakly supervised joint action recognition and evidence localization. For the application of action classification considered here, we assume we are given input videos with action class labels. We use the video representation described above that includes a holistic video representation in addition to a latent region of interest. We aim to learn a classifier that places novel test videos into the correct classes. At the same time, we wish to learn parameters that produce coherent latent regions, regions that are similar across all the training videos. We now provide the details of this algorithm; a summary is provided in Algorithm 1.

Algorithm 1. Training a Similarity Constrained Latent SVM

```
Input : $\mathbf{x} = \{x_1, \ldots, x_N\}$, $\mathbf{y} = \{y_1, \ldots, y_N\}$, $\epsilon$
Output : parameters $w$
Initialize $w_1$
for $t \leftarrow 1$ to $N$ do
    $\{h_1, h_2, \ldots, h_N\}$ = inferLatent ($w_t$, $\mathbf{x}$,$\mathbf{y}$), Sec. 4.2
    $c_{w_t} = \frac{\delta R(w_t)}{\delta w}$, from Eq. 7
    Compute $[w_{t+1}, w_t^*, gap]$, from [21], Alg. 1
    if $gap \leq \epsilon$ or $t == N$ then
        return $w_t^*$
    end if
end for
```

4.1 Learning Formulation

We assume we are given as input a set of training videos $\{x_1, x_2, \ldots, x_n\}$ with action class labels $\{y_1, ..., y_n\}$, $y_i \in Y$, the set of action classes. As in Eq. 1, we aim to learn the parameters w of a scoring function

$$F_w(x, y) = \max_h f_w(x, y, h) \tag{3}$$

that gives a score to labeling a video x with a class label y. This scoring function includes a maximization over latent variables h which encode the latent region of interest within a video. The function $f_w(\cdot)$ contains terms for a holistic bag-of-words video representation in addition to features focused on the latent region of interest.

A standard method for learning parameters to such a model is the latent SVM [1,17] or max-margin hCRF [16]. However, we wish to enforce that the latent regions chosen across all videos of a class are coherent. We argue that this can have two advantages. First, it will result in a model that clusters the latent regions of a video category to produce a discriminatively chosen summary. Second, it acts as an additional regularizer, to smooth out the choice of latent variables. We now provide a formal learning criterion that encompasses this intuition.

The Similarity Constrained Latent SVM chooses model parameters w according to the following criterion:

$$\begin{aligned} &\min_{w,\mathbf{h},\xi>0,\xi^l} ||w||^2 + C_1 \sum_{i=1}^{N} \xi_i + C_2 \sum_{i=1}^{N} \xi_i^l \qquad (4)\\ \text{s.t.}\quad & f_w(x_i, y_i, h_i) - \max_h f_w(x_i, y', h) \geq \Delta(y_i, y') - \xi_i \quad \forall y' \in Y,\ \forall i\\ & \Delta_l(\mathbf{h}, h_i, \mathbf{x}, \mathbf{y}) \leq \xi_i^l \end{aligned}$$

where $\mathbf{h} = \{h_1, ..., h_n\}$, $\mathbf{x} = \{x_1, ..., x_n\}$, and $\mathbf{y} = \{y_1, ..., y_n\}$ are the set of hidden variables, features, and labels for all training videos respectively.

The slack variables ξ_i and corresponding first constraint in Eq. 4 are the usual latent SVM margin constraints on the class labels – model parameters w should correctly classify videos. We use the standard 0/1 loss $\Delta(y_i, y') = \mathbb{1}_{[y_i \neq y']}$ as a penalty on classification error.

The slack variables ξ_i^l and corresponding latter constraint enforce the similarity over latent regions for training videos. We define a loss function $\Delta_l(\mathbf{h}, h_i, \mathbf{x}, \mathbf{y})$ that measures how similar the features within latent region h_i are to those in the remainder of the training images. An individual slack variable ξ_i^l for each training video allows individual videos to be outliers from the set of training videos, though with a penalty controlled by slack-tradeoff parameter C_2.

A variety of loss functions can be used for measuring the similarity of training videos. In our implementation, we choose to measure the distances between pairs of bag-of-words histograms for training videos. We define:

$$\Delta_l(\mathbf{h}, h_i, \mathbf{x}, \mathbf{y}) = \sum_{j=1}^{N} d(h_i, h_j, x_i, x_j) \cdot \mathbb{1}_{[y_i = y_j]} \qquad (5)$$

where $d(h_i, h_j, x_i, x_j)$ is a pairwise dissimilarity measure between two different windows. This dissimilarity is defined as $d(h_i, h_j, x_i, x_j) = -\phi(h_i, x_i)^T \phi(h_j, x_j)$, where $\phi(.)$ is the feature used to represent the hidden regions (bag-of-words histograms in our case). Other variants of loss functions could be used here. For instance, we could also include penalty for similarity between latent regions of videos of different categories. For simplicity, further we use $\Delta_l(\mathbf{h}, h_i)$ instead of $\Delta_l(\mathbf{h}, h_i, \mathbf{x}, \mathbf{y})$ and $d(h_i, h_j)$ instead of $d(h_i, h_j, x_i, x_j)$

In summary, we learn model parameters w that correctly classify videos and enforce a pair-wise similarity of latent regions simultaneously. This enforcement over training videos is different from the standard latent SVM, and presents a novel challenge for learning the parameters w. Next, we describe how we can address this challenge.

4.2 Learning Procedure

We use NRBM, the non-convex bundle optimization by Do and Artières [21] to solve Eq. 4. In a nutshell, the algorithm iteratively builds an increasingly accurate piecewise quadratic approximation to the objective function. During each iteration, a new linear cutting plane is found via a subgradient of the objective function and added to the piecewise quadratic approximation.

Performing inference of the latent variables **h** in each iteration is challenging in the Similarity Constrained Latent SVM model. This is because the addition of a loss function on latent variables $\varDelta_l(\cdot)$ ties the inference of all latent variables together. However, one can use dual decomposition [9], an approximate inference technique, to address this problem.

Computing a Subgradient. The NRBM method operates on $R(w)$, the unconstrained equivalent to Eq. 4. The computation of a subgradient is relatively straight-forward, assuming the inference over **h** can be done.

$$\begin{aligned} R(w) = ||w||^2 + C_1 \sum_{i=1}^{N} \max_{y',h} \Big[f_w(x_i, y', h) + \varDelta(y_i, y') \Big] \\ - \max_{\mathbf{h}} \Big[C_1 \sum_{i=1}^{N} f_w(x_i, y_i, h_i) - C_2 \sum_{i=1}^{N} \varDelta_l(\mathbf{h}, h_i) \Big] \end{aligned} \tag{6}$$

$$\frac{\delta R(w)}{\delta w} = 2w + \sum_{i=1}^{N} \Big[\varPsi(x_i, y_i^*, h_i^*) - \varPsi(x_i, y_i, h_i) \Big] \tag{7}$$

where $y^*, h^*, \mathbf{h} = \{h_i\}$ are defined by:

$$\begin{aligned} (y_i^*, h_i^*) &= \arg\max_{y',h} f_w(x_i, y', h) + \varDelta(y_i, y') \\ \mathbf{h} &= \arg\max_{\mathbf{h}, h_i \in \mathbf{h}} \Big[C_1 \sum_{i=1}^{N} f_w(x_i, y_i, h_i) - C_2 \sum_{i=1}^{N} \varDelta_l(\mathbf{h}, h_i) \Big] \end{aligned} \tag{8}$$

Inferring Latent Variables. At each iteration of NRBM, we must infer the latent variables **h** as specified in Eq. 8. This is similar to finding the *most violated constraint* or performing *loss augmented inference* in training latent / structural SVMs.

This inference is challenging due to the linkage of latent variables h_i across all training videos – the loss $\varDelta_l$ depends on all latent variables. The required inference in Eq. 8 is equivalent to the following:

$$\begin{aligned} &\arg\max_{\mathbf{h}=\{h_i\}} \Big[C_1 \sum_{i=1}^{N} f_w(x_i, y_i, h_i) - C_2 \sum_{i=1}^{N} \sum_{j=1}^{N} \varDelta_l^p(h_i, h_j) \Big], \\ &\varDelta_l^p(h_i, h_j) = d(h_i, h_j) \cdot \mathbb{1}_{[y_i = y_j]} \end{aligned} \tag{9}$$

where $\varDelta_l^p(h_i, h_j)$ is a pairwise loss function on latent variables.

We use dual decomposition [9], an approximate inference technique, to solve this problem. One could use other approximate inference techniques to solve this problem (e.g. LP relaxation, loopy belief propagation), but dual decomposition has two useful properties. It is deterministic, which gives consistency of NRBM (see below), and the structure of our problem lends itself well to the approach of dual decomposition. Dual decomposition breaks the challenging inference task into smaller, tractable ones via the use of auxilliary variables. Additional terms are added into the optimization problem to push the auxilliary variables to agree.

We introduce auxilliary variables $Q = \{q_{ij}\}$ and $\mathbf{h}'$. Each q_{ij} will refer to a latent region for a training video i, and refer to the portion of its loss refering to video j. Terms will be added to enforce that these variables should represent the same latent region all the time. These terms state that $\mathbf{h}' = \mathbf{h}$, $q_{ij} = h'_i, \forall i, j$.

$$\begin{aligned} F = \max_{\mathbf{h}', Q} & \Big[C_1 \sum_{i=1}^{N} f_w(x_i, y_i, h'_i) - C_2 \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta_l^p(q_{ij}, q_{ji})\Big], \\ \text{s.t.} \quad & h'_i = q_{ij} \quad \forall i, j \end{aligned} \tag{10}$$

Hard constraints of this form are difficult to optimize against, and are equivalent to the original, hard problem. In dual decomposition, we relax this and form the Lagrangian:

$$\begin{aligned} L(\boldsymbol{\lambda}, \mathbf{h}', Q) &= C_1 \sum_{i=1}^{N} f_w(x_i, y_i, h'_i) - C_2 \sum_{i=1}^{N} \sum_{j=1}^{N} \Delta_l^p(q_{ij}, q_{ji}) \\ &\quad + \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_{ij}(\hat{h_i}) \big[\mathbb{1}_{[h'_i = \hat{h_i}]} - \mathbb{1}_{[q_{ij} = \hat{h_i}]}\big] \qquad (11) \\ &= \sum_{i=1}^{N} \Big[C_1 f_w(x_i, y_i, h'_i) + \sum_{j=1}^{N} \lambda_{ij}(h'_i)\Big] \\ &\quad + \sum_{i=1}^{N} \sum_{j=1}^{N} \Big[- C_2 \Delta_l^p(q_{ij}, q_{ji}) - \lambda_{ij}(q_{ij})\Big] \qquad (12) \end{aligned}$$

Optimization of this is straight-forward, and involves finding independent variables q_{ij}, h'_i:

$$\begin{aligned} L(\boldsymbol{\lambda}) &= \max_{\mathbf{h}', Q} L(\boldsymbol{\lambda}, \mathbf{h}', Q) \qquad (13) \\ &= \max_{\mathbf{h}'} \sum_{i=1}^{N} \Big[C_1 f_w(x_i, y_i, h'_i) + \sum_{j=1}^{N} \lambda_{ij}(h'_i)\Big] \\ &\quad + \max_{Q} \sum_{i=1}^{N} \sum_{j=1}^{N} \Big[- C_2 \Delta_l^p(q_{ij}, q_{ji}) - \lambda_{ij}(q_{ij})\Big] \qquad (14) \end{aligned}$$

Note that dual decomposition is an *approximate* inference technique. It is not guarranteed to find the maximum setting of the latent variables $\mathbf{h}$. However, this still can be used with the NRBM optimization method. In a nutshell, instead of optimizing $R(w)$, we optimize against an approximation of it. Dual decomposition is deterministic, so NRBM will optimize against this approximation.

4.3 Applying the Model to Test Videos

Given the model parameters w learned using the procedure above, one can use this to perform inference on test videos. This inference procedure will score a video-class label pair, and provide a discriminative latent region for the new test video. We label a test video with class label y^* and latent region h^* as follows:

$$(y^*, h^*) = \arg\max_{y,h} f_w(x, y, h) \tag{15}$$

Brute force enumeration over possible values for y and h for a test video x is feasible, since the set of possible values for these is limited. Y is the set of class labels (tens) and H is the set of objectness clusters (hundreds).

Note that the current application of the model at test time is inductive – the model parameters w learned on the training set can be applied to any unseen test video. It would also be possible to develop a *transductive* variant of this algorithm [8,19] that examines the (unlabelled) test videos together with the training videos in learning. This would enable one to jointly consider the unlabelled test videos and the labelled training videos when choosing model parameters for the discriminative regions of interest.

5 Experiments

In this section we examine the performance of our model for action recognition and evidence localization on the UCF-Sports dataset [22]. The dataset contains 150 videos from 10 action classes: diving, golf swinging, kicking, lifting, horse riding, walking, running, skating, swinging (on the pommel horse and on the floor), and swinging (at the high bar). These videos are taken from real sports broadcasts and the bounding boxes around the subjects are provided for each frame.

Previous results on this dataset use Leave-One-Out Cross Validation (LOO-CV) to report performance. As indicated in Lan et al. [2] parameter tuning (e.g. regularizer weighting) in the LOO-CV scenario is unclear, and choosing the best parameter based on the test dataset performance results in biased evaluation. In addition, some classes of this dataset have significant background correlation between video samples, and LOO-CV may result in memorizing the background rather than learning the action itself. For these reasons, the protocol proposed in Lan et al. [2] is employed to divide the data into training and testing sets.

We compare our algorithm, the Similarity Constrained Latent SVM (SCLSVM), to two baseline algorithms in order to evaluate the contribution of different parts of the proposed model. The first baseline only considers action recognition and the second baseline performs both action recognition and discriminative evidence localization. For all methods, the HOG3D features are extracted via dense sampling and are used to create a 4000 word visual codebook. Additionally, in all experiments, C_1 is set to 100 based on the results of Lan et al. [2]; whereas, C_2 from the proposed method is selected according to the results of cross-validation.

Global bag of words model. We extracted a global bag of words representation for the whole video similar to [23]. This model corresponds to considering only the term $\psi_1(x, y)$ and ignoring $\psi_2(x, h, y)$ in our scoring function f_w. We then trained a multi-class linear SVM classifier upon the global features.

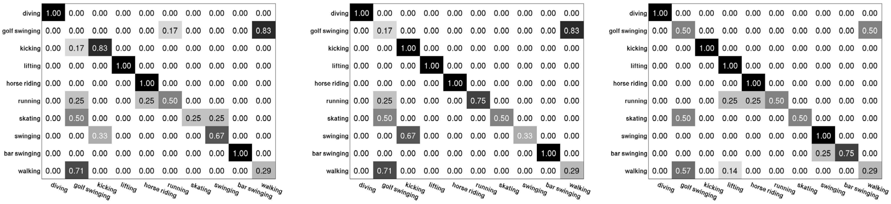
Latent SVM model. This baseline is the same as the model in Eq. 4, except that it does not include a loss term on the dissimilarity of the ROIs between different video samples. In other words, this baseline models the ROI as a latent variable that is set as one of the location hypotheses, and does not penalize the dissimilarity of the ROIs selected for different videos of the same class. This method is implemented according to the general Latent SVM framework [1].

Results analysis. We summarize the numerical results in Table 1, where mean per class recognition accurracy and similarity of the chosen ROIs are compared. First, by analyzing the difference in performance between the Global BoW SVM model and LSVM model, we can conclude that our choice of hypothesis space (utilizing "objectness" operator and mean shift clustering) was reasonable. Second, we can observe that our full model outperforms both baselines (Global BoW SVM and L-SVM) as well as a fully supervised method [2]. Here, we would like to emphasize that the initialization for Latent SVM and SCLSVM is done in the same manner, and use the same domain for the latent variables. SCLSVM outperforms latent SVM due to the modified learning formulation. Confusion matrices for all three models are shown in Fig. 4.

These results demonstrate the positive effect of considering the similarity of discriminative latent regions for videos of the same class. To support this statement, we measured the similarity between the selected ROIs for the test videos and those of the corresponding training videos. For example, when considering the test video with selected region h_t and predicted label y_t, we calculate the normalized similarity between h_t and the ROIs h_j of videos from training set with the same label: $S = \frac{1}{N_t} \sum_{j=1}^{N_t} \phi(h_t, \mathbf{x})^T \phi(h_j, \mathbf{x})$, where N_t is a normalization coefficient, equal to the number of training videos with label y_t. The mean per test video similarity lies within the range $[-1, 1]$. Comparing the similarity for the Latent SVM with that of our model (Table 1, column 2), SCLSVM yields ROIs with greater similarity. As we expected, our model tends to choose regions in test videos that are similar to those used in training.

Table 1. Mean per class accuracy and normalized ROI similarity

Method	Global BoW-SVM	LSVM	**SCLSVM**	Lan et al. [2]
Accuracy	65.4	70.4	**75.3**	73.1
ROI similarity	–	0.1928	**0.2322**	–



(a) Global BoW SVM model (b) Latent SVM model (c) SCLSVM model

Fig. 4. Confusion matrices for different models

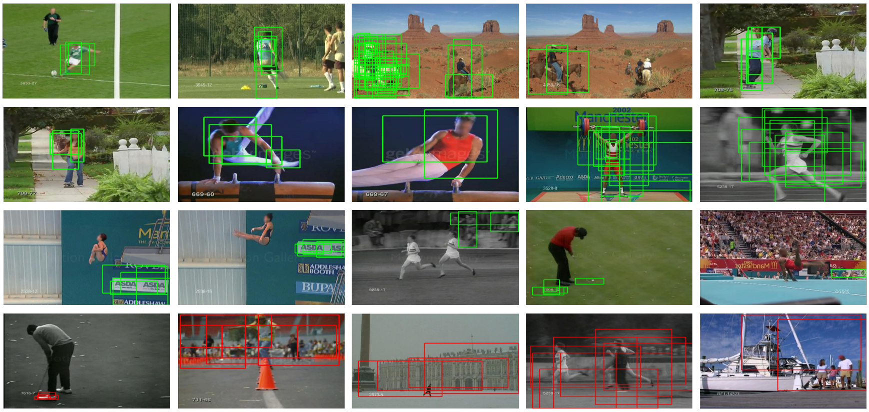


Fig. 5. (Best viewed in color) Visualization of classification results and selected discriminative ROIs indicated by the set of boxes. Green boxes are illustrated for correctly classified videos; red ones for missclassified videos. The first two rows show correctly classified examples with ROIs over the subject of the action. The third row contains correctly classified videos, but with ROIs chosen according to the discriminative context (e.g., videos of the diving action have a distinctive *ASDA* logo in the background). The fourth row illustrates missclassified examples.

Finally, we provide qualitative results produced by the Similarity Constrained Latent SVM model. Fig. 5 demonstrates classification results as well as the localized discriminative evidence (i.e., ROIs) that lead to the classification decision, for several test videos. In many cases (first and second rows of Fig. 5), SCLSVM chooses a ROI that corresponds to the subject of the action (human). However, as shown in the third row of Fig. 5, context is sometimes more distinctive for specific action classes. In this case, the ROI is selected in a way to cover similar contextual regions. For example, the diving class model learns the appearance of the billboard advertisements to classify the diving action, despite the fact that these regions are not semantically related to the action. This outcome is aggravated by the fact that the UCF-Sports dataset contains videos with significant background correlation.

6 Conclusion

We presented a novel learning algorithm, the *Similarity Constrained Latent SVM*, and demonstrated its application to weakly supervised action recognition with discriminative evidence localization. The algorithm learns from training videos with action category labels, and produces a classifier that can label test videos and mark a discriminative region of interest. These regions of interest are learned in a fashion that encourages similar regions to be marked on videos from the same action categories. We demonstrated that the model parameters can be learned efficiently using a combination of NRBM [21] and dual decomposition. Experimental results showed that this approach compares favorably to fully supervised methods.

References

1. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. PAMI 32, 1627–1645 (2010)
2. Lan, T., Wang, Y., Mori, G.: Discriminative figure-centric models for joint action localization and recognition. In: ICCV (2011)
3. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR (2010)
4. Bilen, H., Namboodiri, V., Gool, L.V.: Object and action classification with latent variables. In: Proceedings of the British Machine Vision Conference (2011)
5. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Comput. Surv. 43, 16:1–16:43 (2011)
6. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR (2010)
7. Vezhnevets, A., Ferrari, V., Buhmann, J.M.: Weakly supervised semantic segmentation with a multi-image model. In: ICCV (2011)
8. Joachims, T.: Transductive inference for text classification using support vector machines. In: International Conference on Machine Learning, ICML (1999)
9. Sontag, D., Globerson, A., Jaakkola, T.: Introduction to dual decomposition for inference. In: Sra, S., Nowozin, S., Wright, S.J. (eds.) Optimization for Machine Learning. MIT Press (2011)
10. Komodakis, N.: Efficient training for pairwise or higher order crfs via dual decomposition. In: CVPR, pp. 1841–1848 (2011)
11. Fergus, R., Perona, P., Zisserman, A.: Weakly supervised scale-invariant learning of models for visual recognition. International Journal of Computer Vision 71, 273–303 (2007)
12. Viola, P.A., Platt, J.C., Zhang, C.: Multiple instance boosting for object detection. In: NIPS (2005)
13. Yang, W., Wang, Y., Mori, G.: Recognizing human actions from still images with latent poses. In: CVPR (2010)
14. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (2008)
15. Sorokin, A., Forsyth, D.: Utility data annotation with amazon mechanical turk. In: First IEEE Workshop on Internet Vision (at CVPR). (2008)
16. Wang, Y., Mori, G.: Hidden part models for human action recognition: Probabilistic vs. max-margin. IEEE Trans. PAMI 33, 1310–1323 (2011)
17. Yu, C.N.J., Joachims, T.: Learning structural svms with latent variables. In: ICML (2009)
18. Xu, L., Neufeld, J., Larson, B., Schuurmans, D.: Maximum margin clustering. In: NIPS (2004)
19. Zien, A., Brefeld, U., Scheffer, T.: Transductive support vector machines for structured variables. In: International Conference on Machine Learning, ICML (2007)
20. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC (2008)
21. Do, T.M.T., Artieres, T.: Large margin training for hidden markov models with partially observed states. In: ICML (2009)
22. Rodriguez, M.D., Ahmed, J., Shah, M.: Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR (2008)
23. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC (2009)