OPTIMAL NO-INTERSECTION MULTI-LABEL BINARY LOCALIZATION FOR TIME SERIES USING TOTALLY UNIMODULAR LINEAR PROGRAMMING

Ricardo Cabral[⋆,†]*, João P. Costeira*[⋆]*, Alexandre Bernardino*[⋆]*, Fernando De la Torre*[†]

[⋆]ISR - Instituto Superior Técnico, Lisboa, Portugal
[†]Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

We propose a new model for simultaneously localizing different classes in the same media, casting it as an integer optimization problem. Our model subsumes into a single formulation previous single and multi-class localization methods, as well as allows us to exploit optimal relaxations to the linear domain.

We apply our model to the problem of multi-label multiple instance learning for tagging video collections. Given weakly labeled training samples, where tags for actions in video and objects in images are known but not their locations, our aim is to train classifiers for both detection and localization of said classes on new data. Experimental results demonstrate our approach obtains similar performances when compared to fully supervised methods.

Index Terms— optimization, computer vision, totally unimodular, time series, permutation matrices

1. INTRODUCTION

Object and action recognition have been both a long dream and an outstanding challenge for the computer vision community since its earliest days. These are of utmost importance in enabling applications such as autonomous driving, augmented reality, manufacturing and security. Images arising in these promising scenarios, however, are the main contributors to its complexity: both background clutter as well as extreme viewpoint and scale variations lead to significant intra-class variabilities. Most successful approaches have relied on supervised methods, where the object or action location is provided in training. In the detection phase, a sliding window searches for positive score regions, and a non-maxima suppression step avoids overlapping detections. Sliding window approaches are designed such that the prohibitive space of possible boxes is cleverly searched. Existing approaches employ branch and bound techniques [1], classifier cascades [2] or global context [3]. With many state-of-the-art techniques for recognition following this guideline, significative efforts have been taken to build databases [4] of increasing size to support the training of such methods. We argue, however, that relying on *a priori* localization annotations is not an ideal paradigm towards generalized object and action recognition: as we increase the number of possible categories, manual annotation becomes cumbersome and time consuming.

It has already been shown that classification and localization tasks can benefit from the information of each other. Therefore, we lessen the burden on the labeling process by exploiting weak labels, denoting the presence of a class but not its particular location. Our method looks for associations between regions in both images and video to classes denoting objects or actions. While this topic has been explored in a mutually exclusive single class setting [5], several unaddressed issues surface when performing localization on multiple classes simultaneously. Zha *et al.* [6] formulate a multi-label multiple instance learning framework but explicitly enumerate instances as the result of applying a segmentation algorithm in the images. Vezhnevets and Buhmann [7] instead extend Semantic Texton Learning to a Multitask Multiple Instance Learning framework that performs pixel-wise classification, while regularized by the task of learning geometric context. Similar to previous approaches, we cast the problem in the *Multiple Instance Learning* (MIL) framework, but we learn the classifiers and temporal locations altogether, notwithstanding the massive search space of the latter. We propose a new model for simultaneously localizing different classes in the same video instance, casting it as an integer optimization problem which can be optimally relaxed to a convex optimization problem. Our model's advantage is twofold: first, we subsume into a single formulation previous single and multi-class localization methods [1, 8, 5, 9] and uncover problems not studied before; second, this model enables us to leverage the vast optimization literature to provide off-the-shelf solutions such as LP relaxations or Branch and Cut algorithms.

Support for this research was provided by the Portuguese Foundation for Science and Technology through the Carnegie Mellon Portugal program under the project FCT/CMU/P11. Partially funded by PEst-OE/EEI/LA0009/2013, INCENTIVO/EEI/LA0009/2013, FCT projects Printart PTDC/EEA- CRO/098822/2008, PEst-OE/EEI/LA0009/2013 and the Poeticon++ project from the European FP7 program (grant agreement no. 288382). Fernando De la Torre is partially supported by Grant CPS-0931999 and NSF IIS-1116583. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

2. PREVIOUS WORK

2.1. Efficient Subwindow Search (ESS)

In order to mitigate the complexity of searching the entire space of bounding boxes in an image to find the one that maximizes a classifier score, Lampert *et al.* [1] propose this search is done via a branch and bound technique. By inspecting the dual of an Support Vector Machine (SVM) classifier, they note the resulting weight vector is a linear combination of some training examples. When the feature spaces are histograms to which each pixel or frame contributes toward the count of a single bin, we are able to calculate the score for each element in a picture or movie by replacing it with the respective bin weight. Therefore, we are able to bound the score of all boxes ranging from a maximum and minimum window by summing positive scores of the former and subtracting only the negative ones in the latter. With this in mind, they propose a simple branch and bound search to search the entire space by continuously splitting the space of maximum and minimum boxes in half. Using a priority queue, they first evaluate cases where the upper bound is more promising until the minimum and maximum boxes coincide in a single box with optimal score.

2.2. Multiple Instance Learning (MIL)

Popularized in [10], this class of learning problems extends the typical classification setting to the case where labels are no longer applied individually, but in multi-sets or "bags": a bag is labeled positive if at least one of its instances is positive and negative if none of its constituents are. Nguyen *et al.* [5] cast classification and localization in a SVM MIL formulation where one seeks the sub-window $\varphi(\mathbf{x})$ of $\mathbf{x}$ that maximizes the margin between a positive $\mathbf{d}_i^+$ and negative set $\mathbf{d}_i^-$. The problem is cast as

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \alpha_i, \\ \text{subject to} & \max_{\mathbf{x}\in\mathbf{d}_i^+} \left\{\mathbf{w}^\top\varphi(\mathbf{x}) + b\right\} \geq 1 - \alpha_i, \forall_i \\ & \max_{\mathbf{x}\in\mathbf{d}_i^-} \left\{\mathbf{w}^\top\varphi(\mathbf{x}) + b\right\} \leq -1 + \alpha_i, \forall_i \\ & \alpha_i \geq 0, \end{array} \tag{1}$$

where we want the maximum scoring instance in a negative bag $\mathbf{d}_i^-$ to belong to the negative class and the maximum scoring instance in a positive bag $\mathbf{d}_i^+$ to be on the positive class. Minimization of (1) is therefore an iteration between two steps. First, it fixes the window location and finds classifier parameters $\mathbf{w}, \alpha$. Then, with fixed classifiers, it optimizes over the bounding boxes space $\varphi(\cdot)$ using ESS [1].

Although the SVM in (1) is easily extendable to a multi-class approach, this alone is insufficient when multiple classes co-occur in an image or time series. This Multiple-label MIL problem has only itself recently been studied. Zhou and Zhang [11] map it onto either a multiple class MIL or a standard multi-label problem. The latter and Zhu *et al.* [6] rely on an explicit enumeration of the instances, intractable with our goal of searching the entire space $\varphi(\cdot)$. We argue that although classes are not mutually exclusive in the bag, at an instance-level their co-existence should be penalized. As such, we follow [11] and formulate the multi-label problem by extending (1) to a *one-vs-all* multi-class MIL approach, transferring the multi-label constraints from the classifier to the localization process.

3. MULTI-LABEL LOCALIZATION FOR VIDEO

While translating (1) into a multi-class problem is adequate under the assumption that each pixel is occupied by a single object, it raises the problem of finding the best assignment of segments to all classes simultaneously. This selection model should obey several constraints: 1) select at least N_b segments per labeled class, 2) they should not intersect, 3) segments may have a minimum size s. We consider the time series case with N frames and K classes. Let $\mathbf{X} \in \mathcal{R}^{K\times N}$ be such that $X_{k,n}$ contains the k-th class SVM score in frame n and $\mathbf{B} \in \mathcal{R}^{N\times K}$ a binary matrix whose columns encode frame attributions to a class. Finding the selection that maximizes the joint score can be cast as

$$\begin{array}{lll} \text{maximize} & \text{trace}(\mathbf{XB}) & \text{(2a)} \\ \text{subject to} & \mathbf{1}_N^\top\mathbf{B} \geq s, & \text{(2b)} \\ & \mathbf{B}\mathbf{1}_K \leq 1, & \text{(2c)} \\ & \mathbf{B} \in \mathcal{B}, & \text{(2d)} \\ & \text{card}(\mathbf{B}_k) \geq N_b, & \text{(2e)} \\ & \mathbf{B} \in \{0,1\}, & \text{(2f)} \end{array}$$

where minimum size, no intersection and number of bounding boxes constraints are respectively handled in (2b), (2c) and (2e). We note it is not trivial to codify constraints (2d) through (2f), so we introduce a change to the integral domain $\hat{X}_{k,n} = \sum_{i=1}^{n} X_{k,i}$, where the definition of a bounding box becomes a subtraction of two points. We can obtain $\hat{\mathbf{X}}$ from $\mathbf{X}$ by right multiplying it with an upper triangular matrix of ones $\mathbf{G}$. This allows a natural parameterization of $\mathcal{B}$ using linear constraints, by adding to (2) a constraint imposing causality, *i.e.*, the end of a bounding box comes after its beginning. Problem (2) becomes finding binary matrices $\mathbf{P_e}$, $\mathbf{P_s} \in \mathcal{R}^{N\times K}$, whose nonzero indexes in column k respectively denote ends and starts of N_b boxes for class $k \in K$. We provide a visualization of this change of domain in Fig. 1.

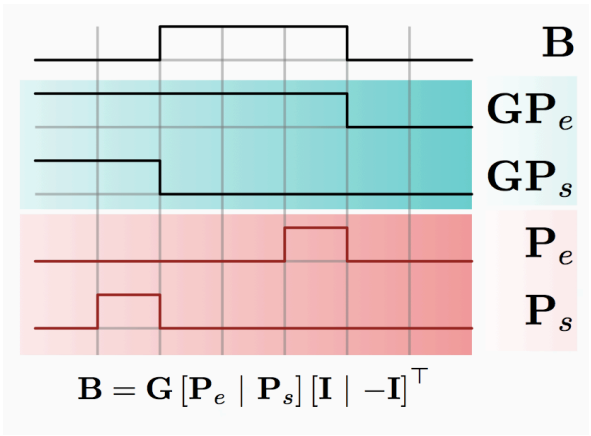

Fig. 1. Visual interpretation of the domain change in (2).

Thus, we rewrite (2) as

$$\begin{aligned}
\text{maximize} \quad & \operatorname{trace}(\mathbf{X}\mathbf{G}\left[\mathbf{P}_e \mid \mathbf{P}_s\right]\left[\mathbf{I} \mid -\mathbf{I}\right]^\top) && \text{(3a)}\\
\text{subject to} \quad & \mathbf{1}_N^\top \mathbf{G}\left[\mathbf{P}_e \mid \mathbf{P}_s\right]\left[\mathbf{I} \mid -\mathbf{I}\right]^\top \geq s, && \text{(3b)}\\
& \mathbf{G}\left[\mathbf{P}_e \mid \mathbf{P}_s\right]\left[\mathbf{I} \mid -\mathbf{I}\right]^\top \mathbf{1}_K \leq 1, && \text{(3c)}\\
& \mathbf{G}\left[\mathbf{P}_e \mid \mathbf{P}_s\right]\left[\mathbf{I} \mid -\mathbf{I}\right]^\top \geq 0, && \text{(3d)}\\
& \mathbf{1}^\top \mathbf{P} \geq N_b, && \text{(3e)}\\
& \mathbf{P}_e, \mathbf{P}_s \in \{0,1\}. && \text{(3f)}
\end{aligned}$$

In formulating our problem as (3), several benefits arise. First, we subsume several works in the same formulation as different combinations of constraints: single class Maximum Subarray Problem algorithms [1, 8] are the solution for (3d) and (3e) as an equality with $N_b = 1$; the Maximum C-disjoint subarray problem [8] generalizes the previous algorithms with $N_b = C$; Nguyen *et al.* [5] assume $K = 1$ and use constraints (3d) and (3e); [9] is the solution for constraints (3d), (3b) (as well as a maximum size) and by forcing a selection of every frame to a class as an equality in (3c) (thus requiring a null class). Additionally, we can use integer programming solvers to tackle the full blown model, as well as exploit optimal linear relaxations of some subproblems.

In this paper, we are interested in finding associations between labels and segments, so we assume N_b is unknown. We explore an optimal relaxation of (3f) to the linear domain that requires a Totally Unimodular constraint matrix. Although (3b) violates this, we note the causality constraint (3d) can enforce length, if we disallow the first and last s integrals as valid ends and starts, by deleting their respective columns from $\mathbf{G}$. The final problem, LP-ESS, becomes

$$\begin{aligned}
\text{maximize} \quad & \operatorname{trace}(\mathbf{X}\left(\mathbf{G}_{:,s+1:N}\mathbf{P}_e - \mathbf{G}_{:,N-s}\mathbf{P}_s\right))\\
\text{subject to} \quad & \left(\mathbf{G}_{:,s+1:N}\mathbf{P}_e - \mathbf{G}_{:,N-s}\mathbf{P}_s\right)\mathbf{1}_K \leq 1,\\
& \mathbf{G}_{s+1:N,s+1:N}\left[\mathbf{P}_e \mid \mathbf{P}_s\right]\left[\mathbf{I}_K \mid -\mathbf{I}_K\right]^\top \geq 0,\\
& 0 \leq \mathbf{P}_e, \mathbf{P}_s \leq 1,
\end{aligned} \qquad (4)$$

where $\mathbf{G}_{a:b,c:d}$ denotes the subset of $\mathbf{G}$ spanning rows and columns from a to b and c to d and we scaled variables $\mathbf{P}_e$,$\mathbf{P}_s$ to the appropriate dimensions. LP-ESS may provide multiple segments per class, each with minimum length s. In this formulation, we do not avoid empty solutions for a subset of the classes. This behavior is beneficial as it increases robustness to the fact that labels may contain errors. To the best of the authors knowledge, we are the first to tackle (4) and the full-blown formulation (3).

Proposition 1 *If a matrix* $\mathbf{A}$ *is Totally Unimodular (TUM), its concatenation with an identity matrix* $\mathbf{I}$ *is also TUM.*

Proof Cf. [12], a necessary and sufficient condition is that for all subsets $\mathcal{R}$ containing rows of $\mathbf{A}$ and $\mathbf{I}$, an assignment $g(r) = \pm 1$ exists such that all elements of the row weighted sum $x = \sum_{r\in\mathcal{R}} g(r)r \in \{-1, 0, 1\}$. Since $\mathbf{A}$ is by definition TUM, we have $\hat{x} = \sum_{r\in\mathcal{R}\setminus\mathcal{I}} g(r)r \in \{-1, 0, 1\}$. Hence, we construct the assignment for the rows of $\mathcal{I}$ by multiplying by the symmetric of $\hat{x}_i$, as each row of $\mathbf{I}$ only has one positive element, each in different columns. ■

Theorem 2 *The constraint matrix of LP-ESS* (4) *is TUM.*

Proof The form of the constraint matrix is obtained by modifying $\mathbf{P}_e$, $\mathbf{P}_s$ in (4) to a vectorized form using Kronecker products $\otimes$. We prove this statement for the matrix arising from $s = 0$, as all others are a subset of this one:

$$\begin{bmatrix} \left[\mathbf{I}_K \mid -\mathbf{I}_K\right] \otimes \mathbf{G}_{1:N,s+1:N} \\ \left[\mathbf{1}_K^\top \mid -\mathbf{1}_K^\top\right] \otimes \mathbf{G}_{1:N,1:N} \\ \mathbf{I}_{N\times 2\times K} \end{bmatrix} \qquad (5)$$

Given Proposition 1, we need only to focus on the first two blocks. These can also be divided in half column-wise, yielding two symmetrical blocks, of the form $\pm\left[\mathbf{I}_K \mid -\mathbf{1}_K\right]^\top$. The Kronecker product of two TUM matrices is TUM as $\det(\left[\mathbf{I}_K \mid -\mathbf{1}_K\right]^\top \otimes \mathbf{G}) = (\det\left[\mathbf{I}_K \mid -\mathbf{1}_K\right]^\top)^K (\det \mathbf{G})^N$. Since $\mathbf{G}$ is an upper triangular matrix, it follows the consecutive ones property and therefore is TUM, therefore we only need to prove matrix $\left[\mathbf{I}_k \mid -\mathbf{1}_K\right]^\top$ is TUM, easily achieved using $g(r) = 1$ for the vector of ones and Proposition 1. ■

4. EXPERIMENTS

In this section, we evaluate the performance of our method in a multi-label weakly-supervised setting. We use the Weizmann dataset (Fig. 2) that consists in 90 sequences of 9 subjects performing 10 actions. We use as features histograms of words obtained by extracting an Euclidean Distance Transform in each frame and clustering in 100 words. We compare our LP-ESS approach to using ESS in the presented MIL framework and to NMSSeg [9], a supervised approach yielding state-of-the-art results on this dataset. To estimate the

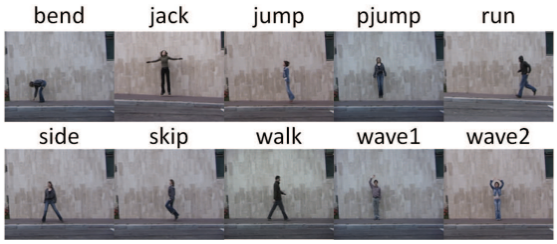

Fig. 2. Typical frames from the Weizmann dataset.

information loss caused by not imposing a minimum number N_b of segments present for the labeled classes, we also present results obtained when using (3) (NumSeg) for localization. We present average frame accuracy, precision and recall results in Table 1 for 20 movies built as concatenations of 10 randomly selected sequences from the dataset. Results show ESS performs poorly in this setting, expectable as it was designed to find a single segment per class. Our method is able to achieve results comparable to state of the art with less label effort. Also, it's worth noting that LP-ESS is not far from the results obtained using the optimal model.

Table 1. 20-test average performance in the Weizmann dataset.

	Accuracy (%)
ESS [1]	16.61 ± 0.6
LP-ESS (4)	78.38 ± 8.3
NumSeg (3)	78.69 ± 8.2
NMSSeg [9]	87.70

5. CONCLUSIONS

We presented a new model for performing simultaneous localization of various classes in both images and video, with no intersection constraints, in the form of an integer program. Experimental results valdate the proposed algorithm as a localization method for performing weakly supervised learning of actions. Although we used this method in the context of visual learning, our theoretical result is potentially applicable to any situation where multiple classes are to be assigned non-intersecting slots of specified minimum duration. this is the case, for instance, of resource allocation in distributed systems, where programs are assigned computation time according to priority scores.

6. REFERENCES

[1] Christoph H. Lampert, Matthew B. Blaschko, and Thomas Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[2] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE ComSoc Conference on Computer Vision and Pattern Recognition*, 2001.

[3] S.J. Hwang and Kristen Grauman, "Reading between the lines: Object localization using implicit cues from image tags," in *Computer Vision and Pattern Recognition*, 2010.

[4] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," *International Journal of Computer Vision*, 2007.

[5] Minh Hoai Nguyen, Lorenzo Torresani, Fernando de la Torre, and Carsten Rother, "Weakly supervised discriminative localization and classification: a joint learning process," *IEEE 12th International Conference on Computer Vision*, 2009.

[6] Zheng-jun Zha, Xian-sheng Hua, Tao Mei, Jingdong Wang, and Guo-jun Qi Zengfu, "Joint multi-label multi-instance learning for image classification," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[7] Alexander Vezhnevets and J.M. Buhmann, "Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning," in *Computer Vision and Pattern Recognition , 2010 IEEE Conference on*, 2010.

[8] S.E. Bae, *Sequential and Parallel Algorithms for the Generalized Maximum Subarray Problem*, Ph.D. thesis, 2007.

[9] Minh Hoai Nguyen, Zhen-zhong Lan, and Fernando De la Torre, "Joint Segmentation and Classification of Human Actions in Video," *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[10] T Dietterich, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, 1997.

[11] Zhi-hua Zhou and M.L. Zhang, "Multi-instance multi-label learning with application to scene classification," *Advances in Neural Information Processing Systems*, 2007.

[12] Alain Ghouila-Houri, "Caractérisation des matrices totalement unimodulaires.," *C. R. Acad. Sci., Paris*, 1962.