# Structural SVM with partial ranking for activity segmentation and classification

Guopeng Zhang and Massimo Piccardi

**Abstract**

Structural SVM is an extension of the support vector machine for the joint prediction of structured labels from multiple measurements. Following a large margin principle, the training of structural SVM ensures that the ground-truth labeling of each sample receives a score higher than that of any other labeling. However, no specific score ranking is imposed among the other labelings. In this paper, we extend the standard constraint set of structural SVM with constraints between "almost-correct" labelings and less desirable ones to obtain a partial-ranking structural SVM (PR-SSVM) approach. Experimental results on action segmentation and classification with two challenging datasets (the TUM Kitchen mocap dataset and the CMU-MMAC video dataset) show that the proposed method achieves better detection and false alarm rates and higher $F1$ scores than both the conventional structural SVM and a comparable unstructured predictor. The proposed method also achieves higher accuracy than the state of the art on these datasets in excess of $14$ and $31$ percentage points, respectively.

**Index Terms**

Structural SVM, Hamming loss, ranking, sequential labeling, hidden Markov model.

## I. INTRODUCTION AND RELATED WORK

Structured prediction addresses the joint assignment of a set of class labels from a set of measurements in the presence of dependencies between the labels. This is a frequent situation with examples ranging from classification of web pages, prediction of protein structure and natural language parsing to segmentation and classification of human activities [1]–[4]. Compared to the separate assignment of single labels, the structured approach is expected to prove more accurate by leveraging the relationships among the labels. The structure is commonly represented in terms of a graphical model, and training and inference algorithms are employed to provide the parametrization of the model and label prediction.

G. Zhang and M. Piccardi are with University of Technology, Sydney (UTS), Australia (e-mail: Guopeng.Zhang@student.uts.edu.au, Massimo.Piccardi@uts.edu.au).

Among the possible training approaches, structural SVM was proposed to extend the large-margin concept of the support vector machine (SVM) to the structured case [5], [6]. It has been applied to a variety of structured tasks with remarkable experimental accuracy [7]–[11]. Training of structural SVM is performed by imposing a pre-determined margin between the score granted to the ground-truth labels and the score granted to any other labeling. Since a predicted labeling may differ from the ground truth to a different extent (from almost correct to completely incorrect), a graded margin such as the Hamming distance is often used. However, while the margin constraint guarantees that the ground-truth labeling receives a higher score than all other labelings, it does not ensure that the other labelings are ranked in correctness order. This may affect applications such as, for instance, human activity segmentation where the manual annotation of the start and end of an activity has a significant degree of uncertainty. In this case, we may wish to ensure that also labelings which are close to the ground truth receive a score higher than other, less qualified labelings. Therefore, the idea proposed in this paper is to augment the constrained optimization of structural SVM with an additional set of constraints ensuring proper scoring of additional, selected labelings. To this aim, we also define a modified Hamming loss to measure the distance between an arbitrary labeling and a predicted labeling. We refer to the proposed technique as partial-ranking structural SVM (PR-SSVM) hereafter.

The task tackled in this paper is the joint segmentation and classification of human activities. In formal terms, we aim to optimally infer a sequence of class labels, $y = \{y_1, \ldots, y_t, \ldots, y_T\} \in Y$, from a given sequence of measurements, $x = \{x_1, \ldots, x_T\}$. We perform classification by detection where $y_t \in \{0, 1\}$, $y_t = 1$ means the presence of an assigned action, and $y_t = 0$ its absence. Following a common model, we assume that the labels are connected in a first-order Markov chain and that each label is connected to the measurement with the same time index. Optimal inference for this model is efficiently provided by dynamic programming algorithms while training is performed by the method described in the following sections. The proposed approach has been tested over two challenging activity sequence datasets: the TUM Kitchen mocap dataset [12] and the CMU-Multimodal Activity video dataset (CMU-MMAC) [13]. The experimental results show that the proposed method achieves an accuracy higher than that of conventional structural SVM and also remarkably higher than previous results.

## II. LOSS FUNCTION AND STRUCTURAL SVM

### A. Loss function

In structural SVM, the margin imposed between the ground-truth labeling, $y^g$, and a predicted labeling, $y$, varies according to a chosen loss function, $\Delta(y^g, y)$, which quantifies the loss carried by a mispredic-

tion. The choice of loss function is typically restricted to functions that decompose over the single labels of a labeling since this facilitates efficient training. The most common choice is the Hamming loss:

$$\Delta_H(y^g, y) = \frac{1}{T}\sum_{t=1}^{T}\delta(y_t^g \neq y_t) \tag{1}$$

where $\delta(\textit{true}) = 1, \delta(\textit{false}) = 0$ and $y_t^g, y_t, t = 1 \dots T$, are the individual labels in labelings $y^g$ and $y$, respectively.

In order to augment structural SVM with an additional set of constraints, a loss function is needed between a reference labeling other than the ground truth and a prediction. We note such a labeling as $\tilde{y}^g$ and the new loss function as $\Delta'(y^g, \tilde{y}^g, y)$. A natural way to define it is as difference of losses with respect to the ground truth:

$$\begin{aligned}\Delta'(y^g, \tilde{y}^g, y) &= \Delta_H(y^g, y) - \Delta_H(y^g, \tilde{y}^g) \\ &= \frac{1}{T}\sum_{t=1}^{T}(\delta(y_t^g \neq y_t) - \delta(y_t^g \neq \tilde{y}_t^g))\end{aligned} \tag{2}$$

In this way, also this loss function remains decomposable over single labels and retains efficient training. Its minimum is a negative value occurring at $\Delta'(y^g, \tilde{y}^g, y = y^g)$, that is, when the prediction is equal to the ground truth. In fact, function $\Delta'(y^g, \tilde{y}^g, y)$ as defined in (2) is a hybrid loss/gain function still rewarding similarity to the ground truth.

### *B. Training by partial ranking*

Given a loss function and a ground-truth labeling, all labelings can be ranked in loss order to form a totally ordered set. In principle, any scoring classifier could be trained not only to assign the highest score to the ground truth, but also to score all labelings in loss order. However, in the structured case the number of distinct labelings is exponential and such an approach would prove infeasible. Therefore, in this work we propose to impose only a partial order relation amongst the labelings by selecting a sub-set to be scored in loss order. We refer to this approach as *partial ranking* for short. While the sub-set can be chosen in any arbitrary way, we argue that selecting labelings which are small perturbations of the ground truth may improve the classifier's accuracy, especially in cases where the ground truth has a degree of uncertainty. In this work, we deal with sequences of binary labels and choose to add only one labeling per sample, $\tilde{y}^g$, obtained by modifying the ground truth by setting to $1$ any $0$ labels preceding and following the ground truth' $1$ labels to account for annotation uncertainty about both the start and the end of a run of positive samples:

$$\tilde{y}_t^g = 1 \text{ if } y_t^g = 1 \text{ or } y_{t-1}^g = 1 \text{ or } y_{t+1}^g = 1 \tag{3}$$

For more general cases with multi-valued labels or non-sequential structures, one can build ground-truth perturbations either randomly or manually. For instance, individual labels could be set to the values that are semantically most similar to the ground truth (such as "orange" for "red" or "adverb" for "adjective"). In general, the additional labelings should be of limited loss with respect to the ground truth.

### C. Structural SVM

Given a set of $N$ training instances, $\{x^i, y^i\}, i = 1 \dots N$, with $y^i$ the ground truth of the $i$-th instance, structural SVM finds a vector of parameters, $w$, by the following constrained minimization:

$$\begin{aligned} &\underset{w,\xi}{\operatorname{argmin}} \|w\|^2 + C\sum_{i=1}^{N} \xi^i \quad s.t. \\ &w^T\phi(x^i, y^i) - w^T\phi(x^i, y) \geq \Delta(y^i, y) - \xi^i, \\ &\quad i = 1 \dots N, \ \forall y \in Y \end{aligned} \tag{4}$$

As in conventional SVM, the objective function aims to limit the error on the training set while at the same time achieving effective generalization. To this aim, term $\sum_{i=1}^{N} \xi^i$ places an upper bound over the total training error, while term $\|w\|^2$ regularizes the solution to encourage generalization. Parameter $C$ is an arbitrary, positive coefficient that balances these two terms. In the constraints, function $\phi(x, y)$ is a feature function that computes structured features from pair $\{x, y\}$ such that $w^T\phi(x, y)$ assigns a score to the pair. The constraint for labeling $y = y^i$ guarantees that $\xi^i \geq 0$. Eventually, $\Delta(y^i, y)$ is the chosen loss function.

The problem in (4) is a quadratic program with linear inequality constraints for which many solvers are available [14], [15]. However, in the structural case the size of $Y$ is exponential and satisfying all the constraints is impossible. For this reason, [6] has proposed a relaxation that can find nearly-correct solutions using only a polynomial-size working set of constraints. The working set is built by searching the sample's most violated constraint at each iteration of the solver:

$$\xi^i = \max_y(-w^T\phi(x^i, y^i) + w^T\phi(x^i, y) + \Delta(y^i, y)) \tag{5}$$

which equates to finding the labeling with the highest sum of score and loss:

$$y^{*i} = \underset{y}{\operatorname{argmax}}(w^T \phi(x^i, y) + \Delta(y^i, y)) \tag{6}$$

This problem is commonly referred to as "loss-augmented inference" due to its resemblance with the common inference. In the case of sequential labels and decomposable losses such as the Hamming loss, it can be efficiently resolved in $O(T)$ time by an appropriately weighted Viterbi algorithm.

## III. EXTENDED PRIMAL PROBLEM

The extension provided by PR-SSVM to the objective function of Eq. (4) consists of the introduction of additional constraints ensuring the score ranking of labelings other than the ground truth. The extended problem is expressed as:

$$\begin{aligned} &\underset{w,\xi,\tilde{\xi}}{\operatorname{argmin}} \|w\|^2 + C \sum_{i=1}^{N} (\xi^i + \tilde{\xi}^i) \quad s.t. \\ &w^T \phi(x^i, y^i) - w^T \phi(x^i, y) \geq \Delta(y^i, y) - \xi^i, \\ &w^T \phi(x^i, \tilde{y}^i) - w^T \phi(x^i, y) \geq \Delta'(y^i, \tilde{y}^i, y) - \tilde{\xi}^i, \\ &i = 1 \ldots N, \ \forall y \in Y \end{aligned} \tag{7}$$

Adding the new constraints brings their total number to $2N\,|Y|$. However, the working-set approach still applies and the loss-augmented inference becomes:

$$\begin{aligned} \tilde{y}^{*i} &= \underset{y}{\operatorname{argmax}}(w^T \phi(x^i, y) + \Delta'(y^i, \tilde{y}^i, y)) \\ &= \underset{y}{\operatorname{argmax}}(w^T \phi(x^i, y) + \Delta(y^i, y) - \Delta(y^i, \tilde{y}^i)) \\ &= \underset{y}{\operatorname{argmax}}(w^T \phi(x^i, y) + \Delta(y^i, y)) \end{aligned} \tag{8}$$

One can see that Eq. (8) is formally identical to Eq. (6) and returns the same labeling, i.e. $\tilde{y}^{*i} \equiv y^{*i}$. However, variable $\tilde{\xi}^i$ is set by the different loss:

$$\tilde{\xi}^i = -w^T \phi(x^i, \tilde{y}^i) + w^T \phi(x^i, y^{*i}) + \Delta'(y^i, \tilde{y}^i, y^{*i}) \tag{9}$$

Eventually, the inclusion of both $\xi$ and $\tilde{\xi}$ in the objective adds up the training loss from both sets of constraints.

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we evaluate the proposed method on two challenging human activity datasets: the TUM Kitchen dataset and the CMU Multimodal Activity (CMU-MMAC) dataset [12], [13]. The TUM Kitchen dataset is a collection of activity sequences recorded in a kitchen equipped with multiple sensors [12]. Four human subjects were asked to set a table using 9 actions, namely *Reaching*, *Carrying*, *TakingSomething*, *LoweringAnObject*, *ReleasingGrasp*, *OpeningADoor*, *ClosingADoor*, *OpeningADrawer*, *ClosingADrawer*. For our experiments, we have used the data from the motion capture sensor for the right and left hands which encode the relevant 3D joints as a 45-D vector. The total number of sequences is 19, each ranging in length between $1,000$ and $6,000$ frames. The CMU-MMAC dataset contains activity sequences from 55 subjects preparing food from various recipes [13]. For our experiments, we have selected the 12 subjects preparing brownies from a dry-mix box, with the activities labeled in 14 classes (see Table IV for the complete list). The videos are from side-view camera *7151062*, with a duration ranging between $8,000$ and $20,000$ frames each.

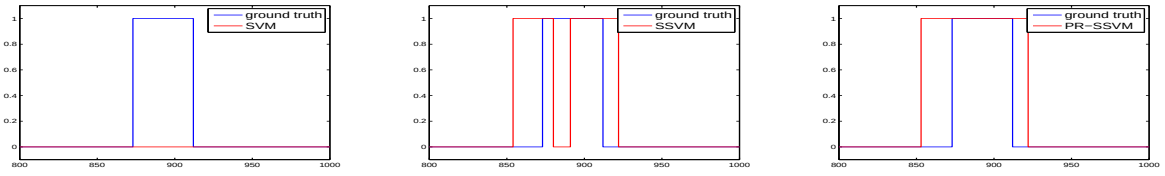
For performance comparison, we have selected the following classifiers: a) as baseline, a standard support vector machine assigning each frame to an activity (*Baseline*); b) a structural SVM classifier using the conventional constraints over the ground-truth labelings (*SSVM*); c) the proposed technique with the augmented set of constraints (*PR-SSVM*). All classifiers were implemented in detector style as a set of binary classifiers, one per activity class. For evaluation, we have recorded performance in terms of detection rate ($DR$), false alarm rate ($FAR$) and $F1$ score. While the detection and false alarm rates describe the trade-off between sensitivity and robustness, the $F1$ score summarizes the performance in a single figure. As parameters, we have tested with $C = [0.01, 100]$ in logarithmic steps, $\epsilon = 0.01$ (default), and a linear, polynomial and RBF kernels for the baseline. The structural SVM classifiers only supports a linear kernel. In all experiments, $C = 0.1$ and the RBF kernel delivered the highest cross-validation accuracy. The software for PR-SSVM ans SSVM was developed using the $SVM^{struct}$ package and its MATLAB wrapper [16], [17], while *libsvm* was used for the baseline [18].

### A. Results on the TUM Kitchen dataset

For TUM Kitchen, we have split the data into a training set with 6 sequences (namely, episodes 1-1 to 1-4, 0-2 and 0-12) and a test set with the remaining 13. Tables I and II report the accuracy at frame level for each class and for the entire test set. These tables show that the tested classifiers achieve very different trade-offs between detection and false alarm rates:

- the baseline reports the lowest $DR$ and $FAR$, implying that most activities will simply go undetected and that its training is biased by the most frequent class (negative);
- structural SVM has a much higher detection rate than the baseline, yet with a rather high $FAR$ (overall, $18.6\%$ for the right hand and $16.8\%$ for the left);
- the proposed technique, PR-SSVM, achieves the best trade-off as it obtains a higher $DR$ than SSVM (overall, $48.0\%$ vs. $45.2\%$ for the right hand, and $37.1\%$ vs. $36.3\%$ for the left hand), together with a lower $FAR$ ($12.7\%$ vs. $18.6\%$ for the right hand, and $9.7\%$ vs. $16.8\%$ for the left hand).

Since it is difficult to rank classifiers based on two rates, we use the $F1$ score for direct comparison. Tables I and II show that PR-SSVM reports the highest overall $F1$ score and for $5$ classes out of $9$ for the right hand, and overall and for $6$ classes out of $9$ for the left one. The overall improvement ranges from $11.6$ to $19.4$ percentage points over SSVM and from $17.7$ to $29.2$ percentage points over the baseline. Fig. 1 shows a typical behavior where a) the baseline misses the activity altogether, (b) SSVM over-segments the activity, while (c) PR-SSVM detects the entire activity as a single segment. Eventually, Table III compares the frame-level accuracy with previous results: although these results cannot be compared directly as the training and test sets differ, the proposed technique shows a remarkable improvement of over $14$ percentage points over the closest result.




(a) (b) (c)

Fig. 1. Example of detection (video 21, right hand, action *ClosingADoor*): a) Baseline; b) Structural SVM (SSVM); c) Partial-ranking structural SVM (PR-SSVM).

### *B. Results on the CMU-MMAC dataset*

For CMU-MMAC, we have divided the $12$ sequences into a training set with the first eight and a test set with the remaining four. As features, we have extracted a dense set of SIFT features from each frame and encoded them as a vector of linearly aggregated descriptors (VLAD) using $k$-means with $32$ clusters [20]. Each measurement results in a $4,096$-D vector. Table IV reports the accuracy results, showing that the relative ranking of the classifiers is unvaried:

- the baseline reports, again, both the lowest $DR$ and $FAR$, with a $DR$ of only $16.1\%$ overall;
- the proposed technique, PR-SSVM, achieves both a higher overall $DR$ ($40.5\%$ vs. $30.0\%$) and lower $FAR$ ($26.9\%$ vs. $41.2\%$) than standard structural SVM;

TABLE I

COMPARISON OF DETECTION RATE, FALSE ALARM RATE AND $F1$ SCORE ON THE TUM KITCHEN DATASET (RIGHT HAND).

| | $DR$ (%) | | | $FAR$ (%) | | | $F1$ score (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| Activity | Baseline | SSVM | PR-SSVM | Baseline | SSVM | PR-SSVM | Baseline | SSVM | PR-SSVM |
| *Reaching* | 0 | 60.9 | 40.2 | 0 | 38.2 | 24.6 | 0 | **23.7** | 21.9 |
| *TakingSomething* | 13.2 | 3.8 | 10.3 | 0.3 | 15.9 | 10.3 | **22.2** | 1.5 | 5.4 |
| *LoweringAnObject* | 0 | 10.3 | 27.8 | 0 | 14.9 | 15.1 | 0 | 5.3 | **13.6** |
| *ReleasingGrasp* | 0 | 16.7 | 18.0 | 0 | 14.4 | 13.4 | 0 | 13.1 | **17.9** |
| *OpeningADoor* | 33.9 | 60.0 | 64.3 | 0.6 | 13.7 | 6.4 | 47.4 | 32.7 | **49.2** |
| *ClosingADoor* | 0 | 16.7 | 48.9 | 0 | 12.4 | 5.4 | 0 | 8.8 | **37.2** |
| *OpeningADrawer* | 0 | 58.9 | 67.1 | 0 | 21.7 | 13.1 | 0 | 19.9 | **31.3** |
| *ClosingADrawer* | 0 | 29.4 | 17.5 | 0 | 11.8 | 8.4 | 0 | **11.5** | 10.1 |
| *Carrying* | 98.0 | 54.2 | 57.9 | 38.7 | 32.0 | 24.7 | **84.2** | 61.3 | 64.3 |
| Overall | 16.1 | 45.2 | 48.0 | 4.4 | 18.6 | 23.7 | 21.3 | 27.4 | **39.0** |

TABLE II

COMPARISON OF DETECTION RATE, FALSE ALARM RATE AND $F1$ SCORE ON THE TUM KITCHEN DATASET (LEFT HAND).

| | $DR$ (%) | | | $FAR$ (%) | | | $F1$ score (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| Activity | Baseline | SSVM | PR-SSVM | Baseline | SSVM | PR-SSVM | Baseline | SSVM | PR-SSVM |
| *Reaching* | 0.2 | 29.5 | 28.9 | 0.0 | 7.6 | 5.9 | 0.4 | 29.2 | **33.4** |
| *TakingSomething* | 50.2 | 81.1 | 68.7 | 0.7 | 29.8 | 10.7 | 9.2 | 28.8 | **52.8** |
| *LoweringAnObject* | 53.5 | 67.7 | 78.0 | 6.3 | 17.7 | 15.9 | **52.6** | 38.5 | 51.3 |
| *ReleasingGrasp* | 0 | 36.4 | 22.5 | 0 | 11.8 | 5.2 | 0 | 21.4 | **26.5** |
| *OpeningADoor* | 0 | 30.3 | 0 | 0 | 32.2 | 14.1 | 0 | **0.2** | 0 |
| *ClosingADoor* | 0 | 0 | 18.5 | 0 | 20.2 | 16.2 | 0 | 0 | **2.4** |
| *OpeningADrawer* | 0 | 0 | 7.8 | 0 | 12.8 | 2.1 | 0 | 0 | **6.2** |
| *ClosingADrawer* | 0 | 12.5 | 37.3 | 0 | 23.5 | 4.9 | 0 | 1.1 | **17.0** |
| *Carrying* | 90.5 | 69.5 | 72.1 | 26.1 | 19.1 | 14.1 | **85.0** | 51.6 | 78.1 |
| Overall | 21.8 | 36.3 | 37.1 | 3.7 | 16.8 | 9.7 | 23.7 | 33.5 | **52.9** |

TABLE III

COMPARISON OF FRAME-LEVEL ACCURACY WITH PREVIOUS RESULTS FOR THE TUM KITCHEN DATASET.

| Method | Average accuracy |
|---|---|
| CRF [12] | 62.8 |
| Switching model [19] | 70.1 |
| Proposed method | 85.0 |

- PR-SSVM reports the highest overall $F1$ score and for 10 classes out of 14, with an overall improvement of 10.9 points over SSVM and 12.2 points over the baseline.

Again, Table V compares the frame-level accuracy with existing results, showing a remarkable improvement of 31 percentage points over the closest value.

TABLE IV

COMPARISON OF DETECTION RATE, FALSE ALARM RATE AND $F1$ SCORE ON THE CMU-MMAC DATASET ("BROWNIES").

| | $DR$ (%) | | | $FAR$ (%) | | | $F1$ score (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| Activity | Baseline | SSVM | PR-SSVM | Baseline | SSVM | PR-SSVM | Baseline | SSVM | PR-SSVM |
| *Closing* | 0 | 45.5 | 54.5 | 12.5 | 44.2 | 40.0 | 0 | 0.6 | **1.0** |
| *Cracking* | 32.5 | 26.6 | 73.4 | 18.2 | 30.5 | 31.2 | **6.5** | 3.5 | 5.3 |
| *None* | 5.8 | 16.8 | 28.4 | 10.3 | 16.7 | 29.4 | 7.5 | 19.5 | **25.1** |
| *Opening* | 16.9 | 43.1 | 45.1 | 13.3 | 30.0 | 33.7 | 9.8 | 12.4 | **15.2** |
| *Pouring* | 18.8 | 75.4 | 70.2 | 25.0 | 70.2 | 59.6 | 14.0 | 26.6 | **27.6** |
| *Putting* | 25.4 | 47.6 | 52.4 | 19.2 | 54.7 | 33.2 | 8.7 | 7.9 | **12.1** |
| *Reading* | 0 | 19.4 | 15.6 | 8.0 | 62.5 | 53.1 | 0 | 0.8 | **3.7** |
| *Spraying* | 7.8 | 3.5 | 14.9 | 24.1 | 10.2 | 22.0 | 1.6 | 1.3 | **1.6** |
| *Stirring* | 71.5 | 10.1 | 29.8 | 70.2 | 10.2 | 29.1 | **35.2** | 14.8 | 29.4 |
| *Switching on* | 15.6 | 24.1 | 44.6 | 25.8 | 33.1 | 23.6 | 3.8 | **4.3** | 4.2 |
| *Taking* | 21.6 | 39.0 | 15.3 | 28.7 | 35.1 | 14.1 | 7.5 | **19.5** | 13.9 |
| *Twisting off* | 0 | 77.5 | 65.9 | 19.5 | 47.6 | 40.0 | 0 | 2.4 | **3.6** |
| *Twisting on* | 0 | 46.4 | 49.8 | 17.3 | 29.3 | 20.2 | 0 | 2.0 | **2.5** |
| *Walking* | 8.8 | 0 | 19.5 | 18.4 | 9.6 | 21.0 | 0.5 | 0 | **8.3** |
| Overall | 16.1 | 30.0 | 40.5 | 22.2 | 41.2 | 26.9 | 7.4 | 8.7 | **19.6** |

TABLE V

COMPARISON OF FRAME-LEVEL ACCURACY WITH PREVIOUS RESULTS FOR THE CMU-MMAC DATASET.

| Method | Average accuracy |
|---|---|
| HMM-MIO [21] | 38.4 |
| CRF [22] | 38.8 |
| Proposed method | 69.8 |

## V. CONCLUSION

In this letter, we have proposed a novel technique for structured prediction enforcing a partial ranking among predicted labelings. This technique is an extension of the versatile structural SVM which joins maximum-margin training with the ability to predict co-dependent labels. The proposed technique, named partial ranking structural SVM (PR-SSVM), imposes a score margin between additional labelings than the ground truth. In particular, in this paper we have enforced a margin between "almost-correct" labelings and the remaining labelings for sequential classification of activities. The results over two contemporary and challenging datasets (TUM Kitchen and CMU-MMAC) show that:

- compared with a baseline classifier providing single-frame classification and standard structural SVM, the proposed PR-SSVM always achieves the highest overall $F1$ scores, with improvements ranging between 11 and 19 percentage points over the runner-up (Tables I, II and IV);
- compared with the other two classifiers, PR-SSVM achieves the most appealing trade-off between $DR$

and $FAR$, with overall values always better than standard structural SVM;

- compared with previous results, PR-SSVM obtains an improvement of over $14$ percentage points on TUM Kitchen and $31$ points on CMU-MMAC (Tables III and V).

In addition, the proposed partial ranking extension is not restricted to sequential classification, but can be applied to any label structure and any sub-set of constraints. Since the proposed loss decomposes over single labels, the efficient loss-augmented inference proper of structural SVM is retained.

## REFERENCES


[1] G. H. Bakir, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan, *Predicting Structured Data*. The MIT Press, 2007.

[2] S. Nowozin and C. H. Lampert, "Structured learning and prediction in computer vision," *Found. Trends. Comput. Graph. Vis.*, vol. 6, no. 3-4, pp. 185–365, Mar. 2011.

[3] N. Smith, "Structured prediction for natural language processing," in *ICML tutorial*, 2009.

[4] M. Hoai, Z.-Z. Lan, and F. De la Torre, "Joint segmentation and classification of human actions in video," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[5] B. Taskar, C. Guestrin, and D. Koller, "Max-margin markov networks," in *NIPS*, 2003.

[6] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *JMLR*, vol. 6, pp. 1453–1484, 2005.

[7] S. B. Wang, A. Quattoni, L.-P. Morency, and D. Demirdjian, "Hidden conditional random fields for gesture recognition," in *Proc. CVPR*, 2006, pp. 2:1521–1527.

[8] U. A.Zien and T.Scheffer, "Transductive support vector machines for structured variables," in *24th Tnternational Conference on Machine Learning*. New York, USA: IEEE, 2007.

[9] X. Zhang and J. Zou, "A structural svm approach for reference parsing," in *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*. Washington D.C., USA: IEEE, 2010.

[10] D. Jurafsky and J. H. Martin, *Speech and language processing*, 2nd ed. New Jersey, USA: Prentice Hall, 2008.

[11] H. Erdogan, "Sequence labeling: Generative and discriminative approaches hidden markov models, conditional random field and structured svm," in *Tutorial at International Conference on Machine Learning and Applications*. Washington D.C., USA: IEEE, 2010.

[12] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition," in *IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), in conjunction with ICCV2009*, 2009.

[13] F. De la Torre, J. K. Hodgins, J. Montano, and S. Valcarcel, "Detailed human data acquisition of kitchen activities: the CMU-multimodal activity database (CMU-MMAC)," in *Workshop on Developing Shared Home Behavior Datasets to Advance HCI and Ubiquitous Computing Research, in conjuction with CHI 2009*, 2009.

[14] L. Bottou and C. Lin, "Support vector machine solvers," *Large-Scale Kernel Machines*, 2007.

[15] O. Chapelle, "Training a support vector machine in the primal," *Neural Computation*, vol. 19, pp. 1155–1178, 2007.

[16] T. Joachims, "$\mathrm{SVM}^{\mathrm{struct}}$: Support vector machine for complex outputs 3.10," 2008. [Online]. Available: http://www.cs.cornell.edu/people/tj/svm_light/svm_struct.html

[17] A. Vedaldi, "A MATLAB wrapper of $\mathrm{SVM}^{\mathrm{struct}}$," 2011. [Online]. Available: http://www.vlfeat.org/~vedaldi/code/svm-struct-matlab.html

[18] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[19] A. Bargi, R. Y. D. Xu, and M. Piccardi, "An online hdp-hmm for joint action segmentation and classification in motion capture data." in *CVPR Workshops*, 2012, pp. 1–7.

[20] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR 2010 - 23rd IEEE Conference on Computer Vision & Pattern Recognition*, 2010, pp. 3304–3311.

[21] E. Zare-Borzeshi, O. Perez-Concha, R. Y. D. Xu, and M. Piccardi, "Joint action segmentation and classification by an extended hidden Markov model," *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1207–1210, 2013.

[22] L. Zhao, X. Wang, G. Sukthankar, and R. Sukthankar, "Motif discovery and feature selection for crf-based activity recognition," in *ICPR*, 2010, pp. 3826–3829.