Appendix

Computational complexity analysis

The computational complexity of SVM-PFS can be easily estimated and compared to SVM-RFE [1], which uses the same feature score but employs backward elimination. SVM-PFS includes $P \cong log(M/N)/log(1-t)$ rounds, and the computationally intensive steps in each round are the SVM-training and weight prediction. When the number of features is large, the main component in SVM training is the computation of the kernel matrix [1] at a cost of L^2M. Weight prediction can be done in $\xi L|F|$, where ξ is the proportion of support vectors in the SVM solution. Putting these together and accumulating the computations of all rounds, the number of computations we get for SVM-PFS is $T_{PFS} = L^2MP + \xi L\left(N-(1-t)M\right)/t$. For $N \gg M$ and large L, i.e. $L \gg N/M$, we have $T_{PFS} \cong L^2MP$. In comparison, SVM-RFE starts by running SVM on the entire set of N features, and hence the number of computations of SVM-RFE is $T_{RFE} \cong \left(N-(1-t)M\right)\left(L^2+\xi L\right)/t$. For $N \gg M$ and large L, we get $T_{RFE} \cong L^2N/t$. Hence, for the typical case where $L \gg N/M$, the theoretical speed gain of SVM-PFS is as follows:

$$\frac{T_{RFE}}{T_{PFS}} \cong \frac{L^2N}{t \cdot L^2MP} = \frac{N\log(1-t)}{t \cdot M\log(M/N)} > \frac{N}{M\log_{(}N/M)} \tag{1}$$

For the case of $L \ll N/M$, a similar analysis shows that PFS has a speed gain of $O(L/\xi)$.

Proofs

Theorem 1:

Proof. Consider adding a new feature $\mathbf{x}^{M+1}$ with a small *fixed* weight w to the classifier. The minimal loss with the new feature is

$$\begin{aligned} &\min_{w_1,..,w_M} \quad L(\textstyle\sum_{j=1}^{M} w_j\mathbf{x}^j + w\mathbf{x}^{M+1}) \\ &= \frac{1}{2}w^2 + \min_{w_1,..,w_M} \tfrac{1}{2}\textstyle\sum_{j=1}^{M} w_j^2 + C\sum_{i=1}^{L}\xi_i \end{aligned} \tag{2}$$

$$\begin{aligned} s.t \quad \forall i \quad & y_i(\textstyle\sum_{j=1}^{M} w_j x_i^j - b) + \xi_i \le 1 - y_i w x_i^{M+1} \\ & \xi_i \ge 0 \end{aligned}$$

This is a perturbed version of the original soft SVM problem over $\boldsymbol{w} = (w_1, ..w_M)$. It has the same optimization argument, and when the constraints are

[1] For online SVM algorithms such as Pegasos [2] the analysis is slightly different and the complexity is $O(f(\epsilon)N)$ where ϵ is the required accuracy. Nevertheless, for large samples, practical ϵ is small, $f(\epsilon)$ is typically larger than L, and the speedup factor of SVM-PFS is retained.

written in the form of $\mathbf{Av} \geq \mathbf{B}$, the perturbation only occurs in $\mathbf{B}$. Specifically, denote $\hat{\mathbf{x}}^{M+1} = \mathbf{y} \otimes \mathbf{x}^{M+1}$ and by $\mathbf{X}$ the data matrix of the first M features (i.e. $\mathbf{X}_{ij} = x_i^j$). The original SVM problem has

$$\mathbf{A} = \begin{pmatrix} -diag(\mathbf{y})\mathbf{X} & -\mathbf{I} \\ 0 & \mathbf{I} \end{pmatrix}$$
$$\mathbf{v} = \begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{\xi} \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} -\mathbf{1} - b\mathbf{y} \\ 0 \end{pmatrix}$$

and the perturbed version has $\mathbf{B} = [-\mathbf{1} - b\mathbf{y} + w\hat{\mathbf{x}}^{M+1}, 0]^T \in \mathbb{R}^{2L}$. Consider the problem's value function $V(\mathbf{A}, \mathbf{B})$ which maps a matrix×vector $(\mathbf{A}, \mathbf{B})$ parameters to the optimal value obtained with these parameters. Theorem 14.2 in [3] gives existence conditions and a general formula for the gradient of this function. Specifically it implies that if both the primal and dual solution are unique at $(\mathbf{A}, \mathbf{B})$, then V is differentiable at this point and its gradient w.r.t changes in $\mathbf{B}$ (for constant $\mathbf{A}$) is the vector of Lagrange multipliers achieving the dual's optimum. For our problem formulation as $\mathbf{Av} \geq \mathbf{B}$, this vector is $[-\boldsymbol{\alpha}, \boldsymbol{\eta}]$. Rewriting Eq. 2 using a Taylor approximation we therefore get

$$\begin{aligned} &\frac{1}{2}w^2 \quad + \quad V(\mathbf{A}, \mathbf{B} + [w\hat{\mathbf{x}}^{M+1}, 0]) \\ &\quad = \tfrac{1}{2}w^2 + V(\mathbf{A}, \mathbf{B}) + [-\boldsymbol{\alpha}, \boldsymbol{\eta}]^T [w\hat{\mathbf{x}}^{M+1}, 0] + O(w^2) \end{aligned} \tag{3}$$

For small enough w the $O(w^2)$ terms are negligible and minimizing this expression w.r.t choice of $\mathbf{x}^{M+1}$ amounts to minimization of $[-\boldsymbol{\alpha}, \boldsymbol{\eta}]^T [w\hat{\mathbf{x}}^{M+1}, 0] = -w\boldsymbol{\alpha} \cdot \hat{\mathbf{x}}^{M+1}$. This is equivalent to maximization of $(\boldsymbol{\alpha} \cdot \hat{\mathbf{x}}^{M+1})^2$ and choosing $sign(w) = sign(\boldsymbol{\alpha} \cdot \hat{\mathbf{x}}^{M+1})$.

Theorem 2:

Proof. For simplicity the analysis is described for hard SVM (i.e., without the soft margin term $C\sum_{i=1}^{L}\xi_i$), and without a bias term b in the classifier. However, it can be easily extended to soft SVM. Extending this result to SVM with a bias term is not straightforward using the techniques introduced here. However, they are directly applicable when the bias term is replaced by adding a constant feature column.

Denote by $\boldsymbol{\alpha}_{sv}$ the reduction of $\boldsymbol{\alpha}^{old}$ to coordinates SV. Then $\boldsymbol{\alpha}_{sv}$ is a maximum of the old dual problem reduced to the variables in SV, i.e. $\boldsymbol{\alpha}_{sv} = \arg\max_{\mathbf{v} \geq 0} \mathbf{v} \cdot \mathbf{1} - \frac{1}{2}\mathbf{v}^T \hat{\mathbf{K}}_{sv}\mathbf{v}$. Since $\boldsymbol{\alpha}_{sv} > 0$, it is an interior maximum. The function $\mathbf{v} \cdot \mathbf{1} - \frac{1}{2}\mathbf{v}^T \hat{\mathbf{K}}_{sv}\mathbf{v}$ is concave and has a single interior maximum. Hence $\boldsymbol{\alpha}_{sv}$ is this maximum and can be found by equating the gradient to 0.

$$\mathbf{1} - \hat{\mathbf{K}}_{sv}\boldsymbol{\alpha}_{sv} = \mathbf{0} \rightarrow \boldsymbol{\alpha}_{sv} = \hat{\mathbf{K}}_{sv}^{-1}\mathbf{1}.$$

The predicted weight for feature $\mathbf{x}^{M+1}$ is $w^{pred} = \boldsymbol{\alpha}^{old} \cdot \hat{\mathbf{x}}^{M+1} = \boldsymbol{\alpha}_{sv} \cdot \mathbf{u} = \mathbf{1}^T \hat{\mathbf{K}}_{sv}^{-1}\mathbf{u}$. Under the assumption that SV did not change when feature

$\mathbf{x}^{M+1}$ was added, the new $\boldsymbol{\alpha}$ vector is given by $(\hat{\mathbf{K}}_{sv}^{new})^{-1}\mathbf{1}$, where $\hat{\mathbf{K}}^{new} = \hat{\mathbf{K}} + \hat{\mathbf{x}}^{M+1}(\hat{\mathbf{x}}^{M+1})^T$ and $\hat{\mathbf{K}}_{sv}^{new} = \hat{\mathbf{K}}_{sv} + \mathbf{u}\mathbf{u}^T$ is the SV-sub-matrix of $\hat{\mathbf{K}}^{new}$. The weight of the new feature is hence $w^{real} = \mathbf{1}^T(\hat{\mathbf{K}}_{sv} + \mathbf{u}\mathbf{u}^T)^{-1}\mathbf{u}$. We get

$$\frac{w^{pred}}{w^{real}} = \frac{\mathbf{1}^T\hat{\mathbf{K}}_{sv}^{-1}\mathbf{u}}{\mathbf{1}^T(\hat{\mathbf{K}}_{sv} + \mathbf{u}\mathbf{u}^T)^{-1}\mathbf{u}} = \frac{\mathbf{1}^T\hat{\mathbf{K}}_{sv}^{-1}\mathbf{u}}{\mathbf{1}^T\left(\hat{\mathbf{K}}_{sv}^{-1} - \frac{\hat{\mathbf{K}}_{sv}^{-1}\mathbf{u}\mathbf{u}^T\hat{\mathbf{K}}_{sv}^{-1}}{1+\mathbf{u}^T\hat{\mathbf{K}}_{sv}^{-1}\mathbf{u}}\right)\mathbf{u}} = \mathbf{1} + \mathbf{u}^T\hat{\mathbf{K}}_{sv}^{-1}\mathbf{u}$$

where we used *Woodbury's identity* for $(\hat{\mathbf{K}}_{sv} + \mathbf{u}\mathbf{u}^T)^{-1}$. Since $\hat{\mathbf{K}}_{sv}$ is positive semi-definite , $\mathbf{u}^T\hat{\mathbf{K}}_{sv}^{-1}\mathbf{u} \geq 0$, from which the left inequality of the theorem follows. The left inequality follows from the inequality $\mathbf{u}^T\hat{\mathbf{K}}_{sv}^{-1}\mathbf{u} \leq ||\mathbf{u}||^2/\lambda_{min}(\hat{\mathbf{K}}_{sv})$.

Theorem 3:

Proof. The optimal $\boldsymbol{\alpha}$ vector is a solution of a constrained maximization problem:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^L} \boldsymbol{\alpha}\cdot\mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^T\hat{\mathbf{K}}\boldsymbol{\alpha} \; s.t. \; 0 \leq \boldsymbol{\alpha} \leq C$$

Where $\hat{\mathbf{K}}$ here is the signed Gram Matrix $\hat{\mathbf{K}}_{ij} = y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$. However, we can solve the maximization problem without the constraints, and if the solution found obeys the constraints then it is clearly also the solution of the constrained problem. The unconstrained solution can be found by plain differentiation to be $\boldsymbol{\alpha} = \hat{\mathbf{K}}^{-1}\mathbf{1}$, where the matrix $\hat{\mathbf{K}}$ is invertible since $M \gg L$, and the features are sampled from a continuous distribution, hence they are in general position.

We will now analyze the behavior of the Gram matrix entrees. For simplicity of notation, we will drop the hat $\hat{\cdot}$ from $\hat{\mathbf{K}}$ in the following.

The diagonal entrees: For all $1 \leq i \leq L$, $\mathbf{K}_{ii} = \sum_{k=1}^{M}(x_i^k)^2$. As a sum of many independent variables, this quantity has a normal distribution. We have $E[\mathbf{K}_{ii}] = \sum_{k=1}^{M} E(x_i^k)^2 = M$. $E[\mathbf{K}_{ii}^2]$ is given by

$$E[\mathbf{K}_{ii}^2] = E[(\sum_{k=1}^{M}(x_i^k)^2)^2] = E[\sum_{k=1}^{M}\sum_{k'=1}^{M}(x_i^k)^2(x_i^{k'})^2]$$

$$= E[\sum_{k=1}^{M}(x_i^k)^4 + \sum_{k\neq k'}(x_i^k)^2(x_i^{k'})^2] = MJ + M(M-1)$$

The variance is hence given by $E[\mathbf{K}_{ii}^2] - (E[\mathbf{K}_{ii}])^2 = MJ + M^2 - M - M^2 = M(J-1)$. This can be summarized by stating that for all i, $\mathbf{K}_{ii} \sim N(M, \sqrt{M(J-1)})$.

Off diagonal entrees: Again $\mathbf{K}_{ij} = \sum_{k=1}^{M} x_i^k x_j^k$ is normally distributed as a sum of many independent terms. For $i \neq j$ we have $E[\mathbf{K}_{ij}] = E[\sum_{k=1}^{M} x_i^k x_j^k] = 0$ since x_i^k, x_j^k are independent for all k. The second moment, which is equal to the variance (as $E\mathbf{K}_{ij} = 0$) is given by

$$E[\mathbf{K}_{ij}^2] = E[\sum_{k=1}^{M}(x_i^k)^2(x_j^k)^2 + \sum_{k\neq k'} x_i^k x_j^k x_i^{k'} x_j^{k'}] = M$$

Hence $\mathbf{K}_{ij} \sim N(0, \sqrt{M})$ for all $i \neq j$.

In addition, for every $1 \leq i, j, l, m \leq L$ we have that $\mathbf{K}_{ij}, \mathbf{K}_{lm}$ are uncorrelated Gaussians, hence independent. This is trivial for $i \neq l, j \neq m$ due to the independence of the underlying variables, but it also hold if one of these is not true. For example if $j = m$:

$$E[\mathbf{K}_{im}\mathbf{K}_{lm}] = E[\sum_{k=1}^{M} x_i^k x_m^k \sum_{k'=1}^{M} x_l^{k'} x_m^{k'}]$$

$$= E[\sum_{k=1}^{M} x_i^k x_l^k (x_m^k)^2 + \sum_{k\neq k'} x_i^k x_l^{k'} x_m^{k'} x_m^k] = 0$$

In a similar way we can show the Independence of two matrix elements of the form K_{ii}, K_{ij}.

Based on the above analysis, we can write $\mathbf{K} = M\mathbf{I} + \sqrt{M}\mathbf{W}$, where $\mathbf{W}$ is an $L \times L$ matrix with independent elements

$$\mathbf{W}(i,i) \sim N(0, \sqrt{J-1})$$
$$\mathbf{W}(i,j) \sim N(0,1) \;\; for\, i \neq j$$

Writing $\mathbf{K} = M(\mathbf{I} + \frac{1}{\sqrt{M}}\mathbf{W})$, and assuming large M, $\mathbf{K}^{-1}$ is approximately given by $\mathbf{K}^{-1} = \frac{1}{M}(\mathbf{I} - \frac{1}{\sqrt{M}}\mathbf{W})$, i.e. this is true up to terms of magnitude $\frac{1}{M}\mathbf{W}$ which are negligible for large M. We can now solve for $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{1} = \frac{1}{M}(\mathbf{I} - \frac{1}{\sqrt{M}}\mathbf{W}) \cdot \mathbf{1} = \frac{1}{M}(\mathbf{1} - \frac{\mathbf{1}}{\sqrt{\mathbf{M}}}\mathbf{W} \cdot \mathbf{1}) \tag{4}$$

Consider coordinate i of $\mathbf{W} \cdot \mathbf{1}$, $[\mathbf{W} \cdot \mathbf{1}]_i = \sum_{i\neq j} \mathbf{W}_{ij} + \mathbf{W}_{ii}$. This variable is normal since it is a sum of normal variables. We have $E[\mathbf{W} \cdot \mathbf{1}]_i = 0$ and from the independence of $\mathbf{W}$'s elements we have $var(\mathbf{W} \cdot \mathbf{1}_i) = E[(\sum_{i\neq j} \mathbf{W}_{ij} + \mathbf{W}_{ii})^2] = \sum_{j\neq i} E[\mathbf{W}_{ij}^2] + E[\mathbf{W}_{ii}^2] = L - 1 + J - 1 = L + J - 2$. Hence $[\mathbf{W} \cdot \mathbf{1}]_i \sim N(0, \sqrt{L + J - 2})$.

From Eq. 4 and these considerations we get for all i that $\alpha_i = \frac{1}{M}(1 - \frac{1}{\sqrt{M}}[\mathbf{W} \cdot \mathbf{1}]_i = \frac{1}{M}(1 + \frac{\sqrt{L+J-2}}{\sqrt{M}}\xi_i)$ where we define $\xi_i = -[\mathbf{W} \cdot \mathbf{1}]_i / \sqrt{(L + J - 2)}$.

Clearly for large enough M, α_i found here obey the constraints $0 \leq \boldsymbol{\alpha} \leq C$, and hence these are the solution of the SVM dual problem.

We now express the PFS score $h(\mathbf{x})$ using the formula for $\boldsymbol{\alpha}$. For each feature vector $\mathbf{x}$ we have:

$$h(\mathbf{x}) = \sum_{i=1}^{L} \alpha_i y_i x_i = \tfrac{1}{M}(\textstyle\sum_{i=1}^{L} y_i x_i + \sqrt{\dfrac{L+J-2}{M}} \displaystyle\sum_{i=1}^{L} y_i x_i \xi_i) \tag{5}$$

Now each term in the last sum is a Gaussian $y_i x_i \xi_i \sim N(0, x_i^2)$ and their sum $\sum_{i=1}^{L} y_i x_i \xi_i$ is a Gaussian with zero mean and variance $\sum_{i=1}^{L} x_i^2 = ||\mathbf{x}||_2^2$. We can therefore define $\xi_h = \sum_{i=1}^{L} y_i x_i \xi_i / ||\mathbf{x}||_2$, and $\xi_h \sim N(0,1)$.

Since $E[\mathbf{x}] = \frac{1}{L}\sum_{i=1}^{L} x_i = 0$ and $var(\mathbf{x}) = E[x^2] = \frac{1}{L}\sum x_i^2 = \frac{||\mathbf{x}||_2^2}{L}$ we have

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\frac{1}{L}\sum_{i=1}^{L} y_i x_i - (\frac{1}{L}\sum_{i=1}^{L} x_i)\cdot(\frac{1}{L}\sum_{i=1}^{L} y_i)}{\sigma(\mathbf{x})\sigma(\mathbf{y})} = \frac{\sum_{i=1}^{L} y_i x_i}{\sqrt{L}\sigma(\mathbf{y}) \cdot ||\mathbf{x}||_2} \tag{6}$$

We can now get the result by isolating $\sum_{i=1}^{L} y_i x_i$ from Eq. 6 and putting it into Eq. 5:

$$\begin{aligned} h(\mathbf{x}) &= \frac{1}{M}(\,\rho(\mathbf{x},\mathbf{y})\sqrt{L}\sigma(\mathbf{y})||\mathbf{x}||_2 + \sqrt{\frac{L+J-2}{M}}||\mathbf{x}||_2\xi_h) \\ &= \frac{\sqrt{L}\sigma(\mathbf{y})||\mathbf{x}||_2}{M}(\rho(\mathbf{x},\mathbf{y}) + \sqrt{\frac{L+J-2}{LM\sigma^2(\mathbf{y})}}\xi_h) \end{aligned}$$

Additional results

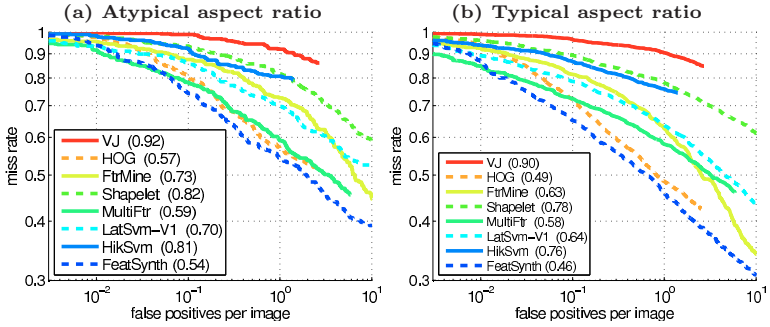

Fig. 1. Additional results on the Caltech pedestrian training dataset partitioned by aspect ratio. Left: Results on the subset of Caltech training dataset including pedestrians with atypical aspect ratios. **Right:** Typical aspect ratios. See [4] for partition and compared method details.

References

1. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46** (2002)

2. Shalev-Shwartz, S., Singer, Y., Srebro, N.: Pegasos: Primal estimated sub-gradient solver for svm. In: International Conference on Machine Learning (ICML). (2007)
3. G. M. Lee, N.T., Yen, N.: Quadratic programming and affine variational inequalities. Springer Netherlands (2005)
4. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: CVPR. (2009)