Novel Fusion Methods for Pattern Recognition

Muhammad Awais,
Fei Yan, Krystian Mikolajczyk and Josef Kittler

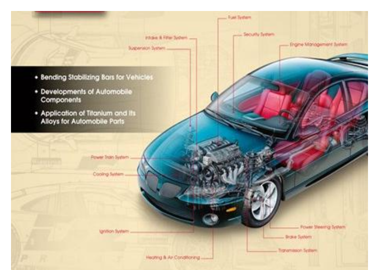
ECML-PKDD 2011

Motivation

- ➢ Automatic analysis of visual information

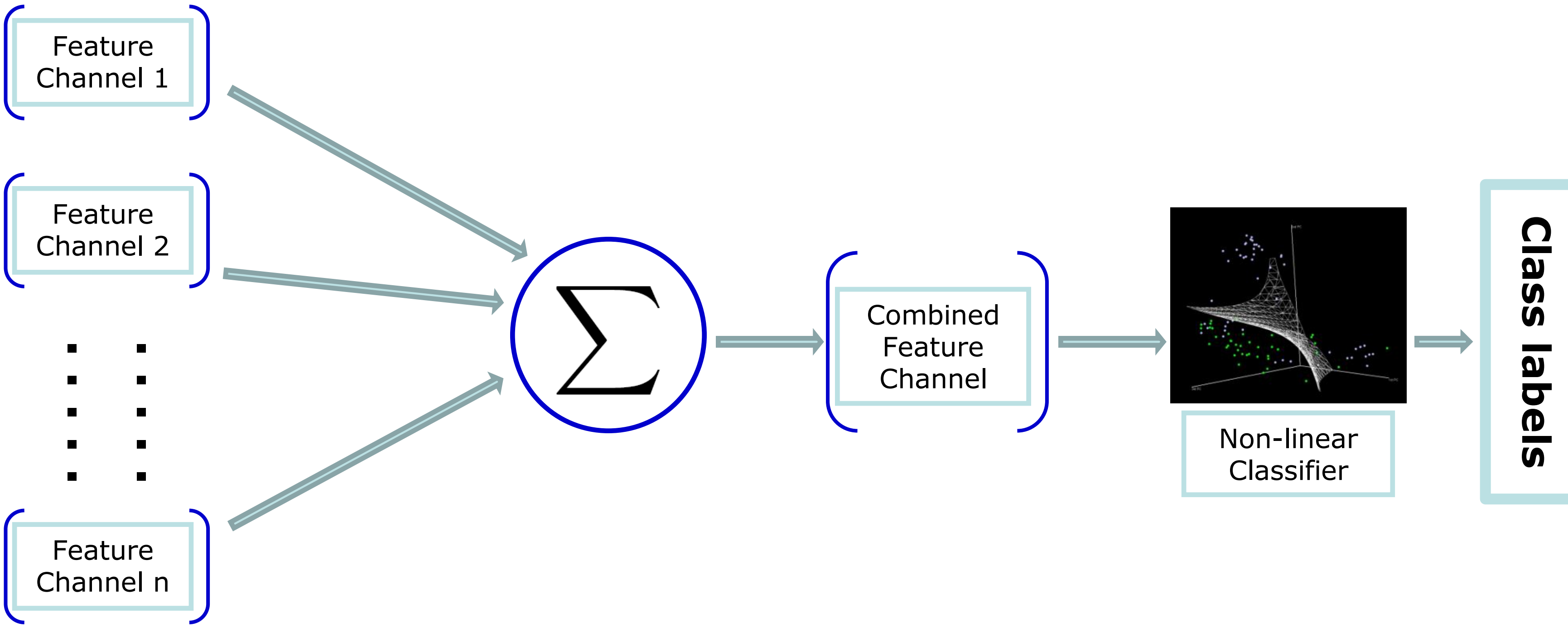
Crime prevention Visually impaired multimedia documents Image content search

- ➢Classification of photos



news agencies art catalogues industrial components trademarks

Problem Statement

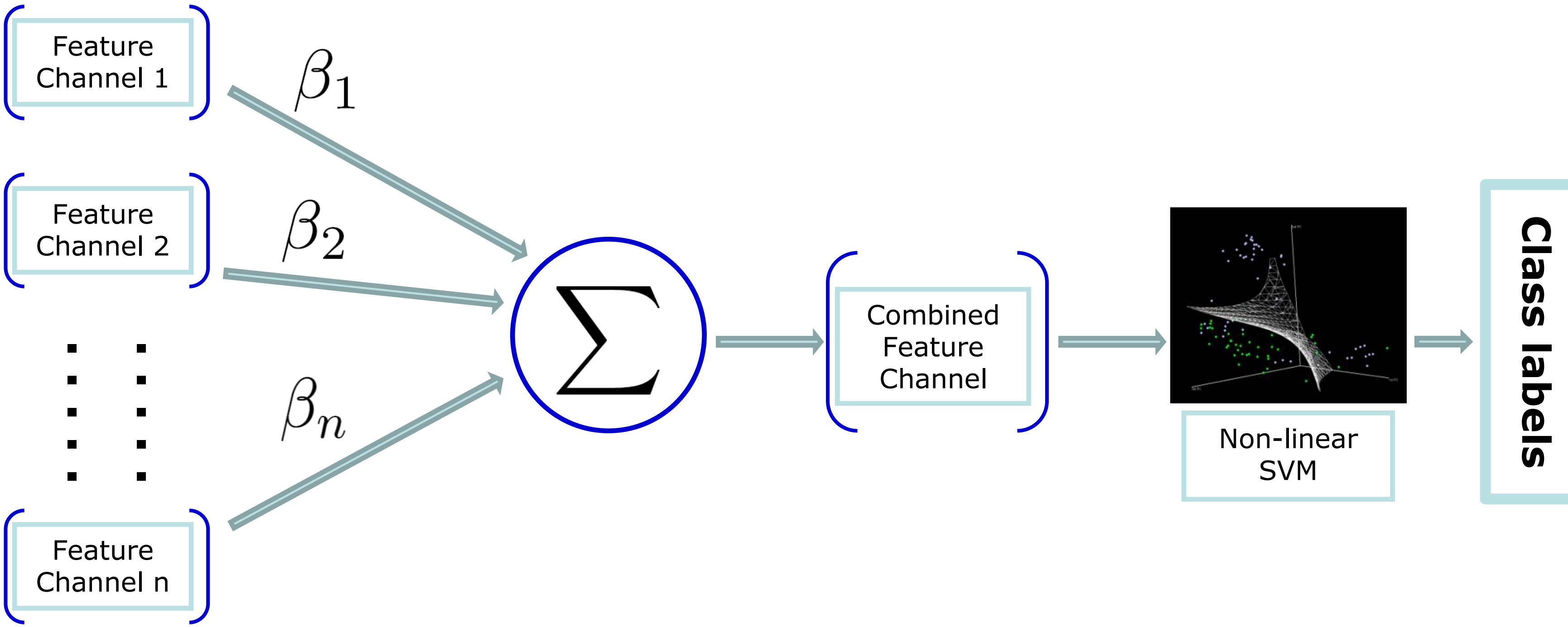


- Given a set of n features channels (kernels). Aim is to find the optimal way of combining these features channels.

Contents

- Existing fusion techniques.
- Proposed Classifier fusion techniques.
  - Binary CLF.
  - NLP-νMC.
  - NLP-β.
  - NLP-B.
- Extended stacking.
- Evaluation on challenging datasets.
- Conclusions.

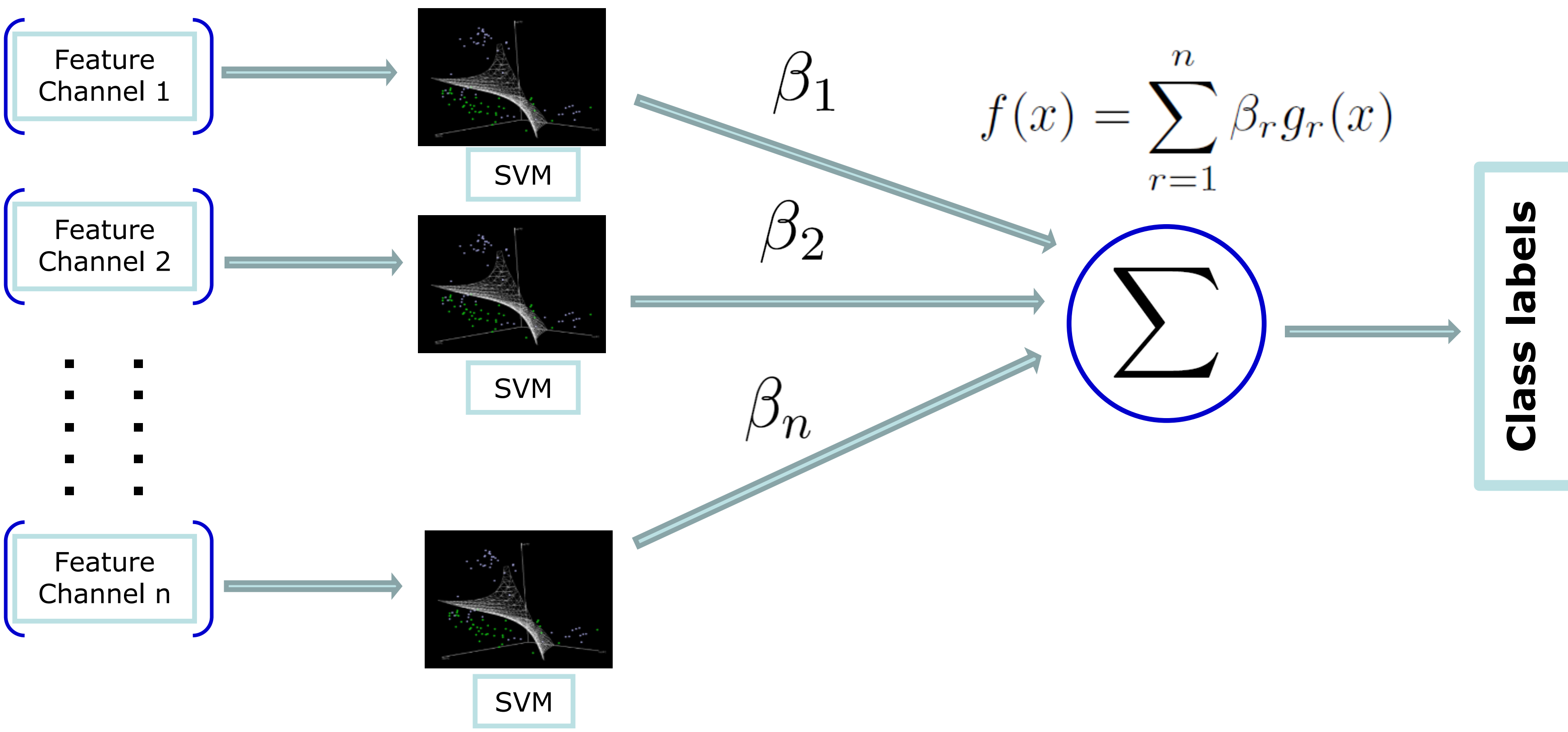
Multiple Kernel Learning (MKL)



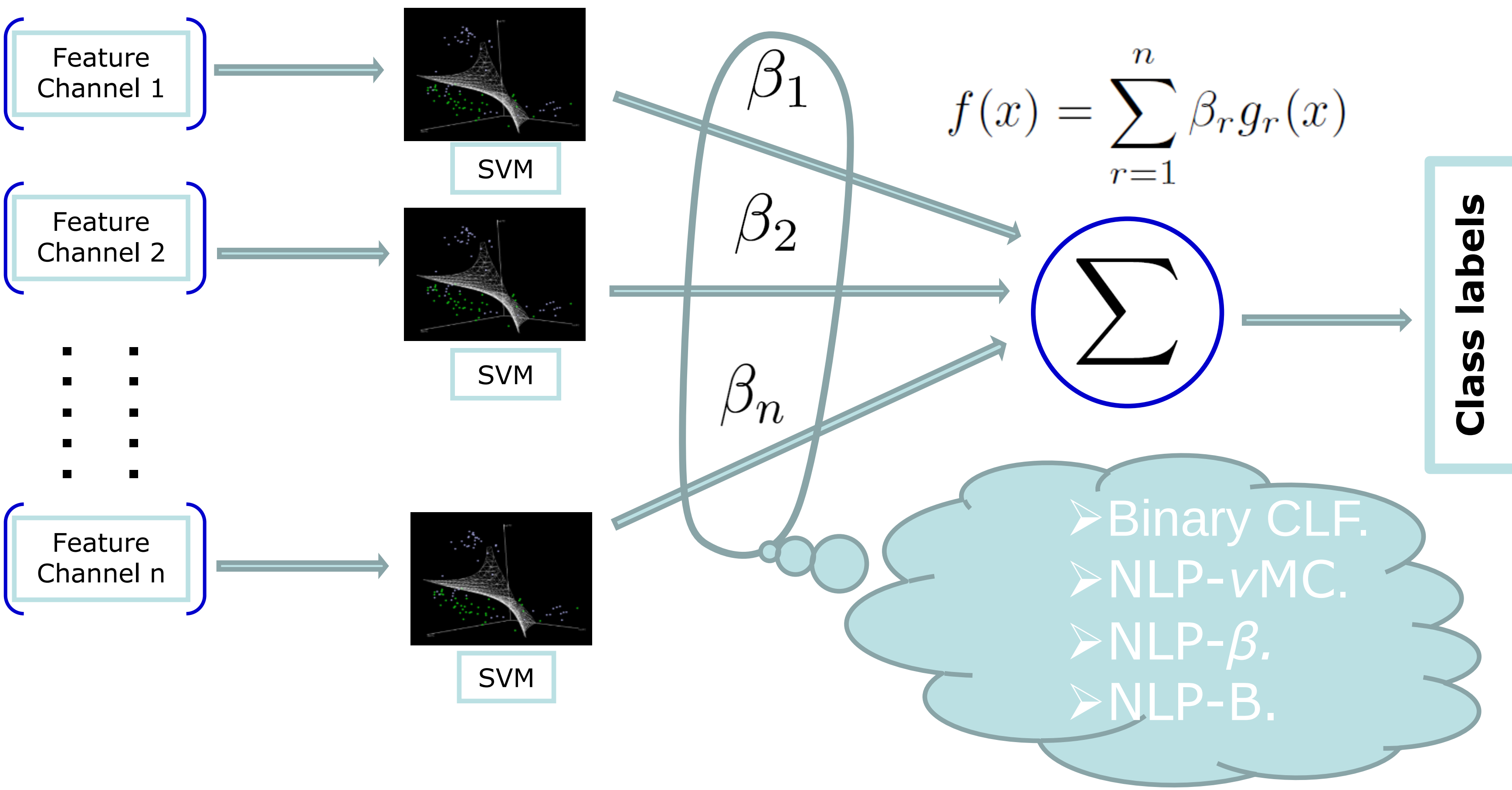
- MKL maximizes the soft margin to obtain optimal weights, for the convex combination of base kernels.

$$K = \sum_{p=1}^{n} \beta_p K_p$$

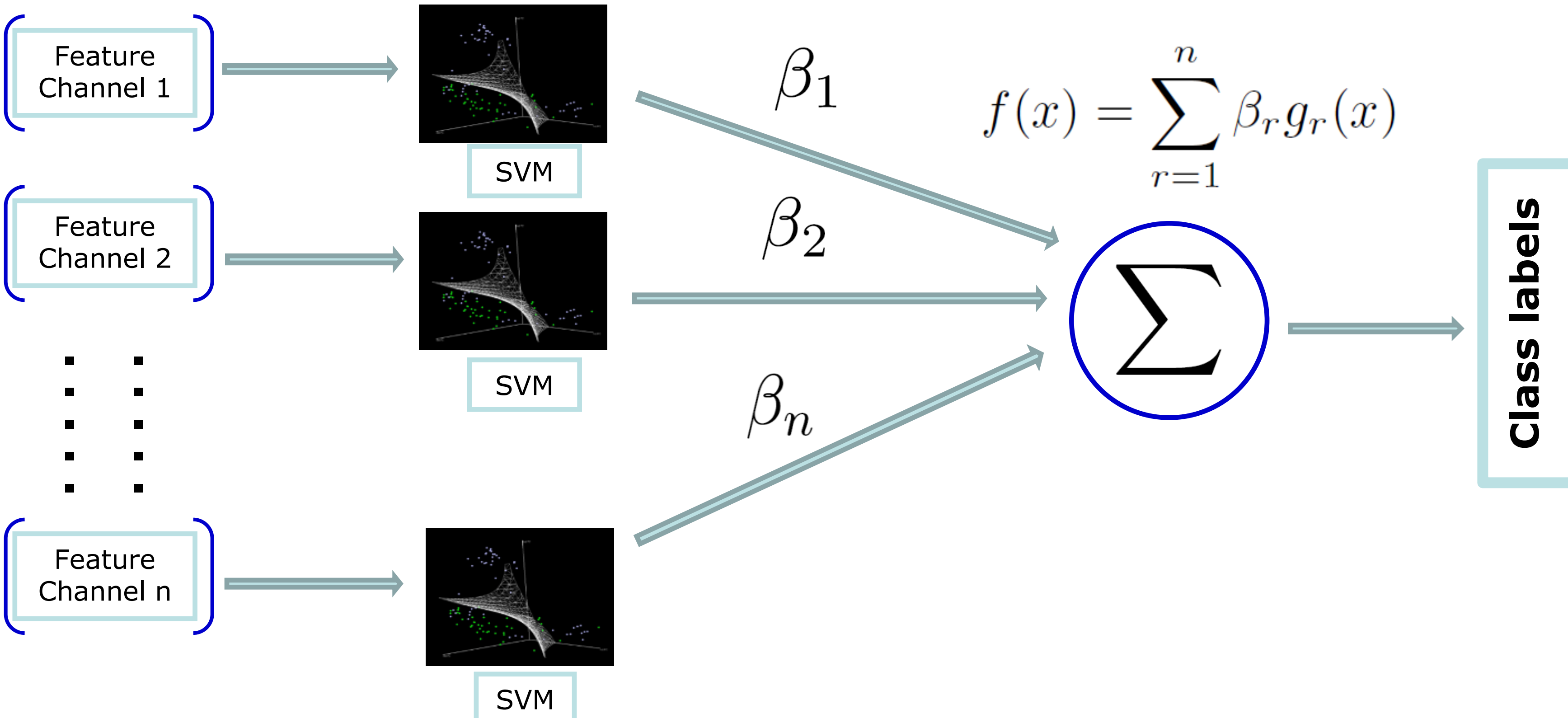
Classifier Level Fusion (CLF)



Classifier Level Fusion (CLF)



Classifier Level Fusion (CLF)



$$\rho := \min_{1 \leq i \leq m} y_i f(x_i) = \min_{1 \leq i \leq m} y_i \sum_{r=1}^{n} \beta_r g_r(x_i)$$

Binary CLF with Non-Linear Constraints

- Extension of *ν*–LP-AdaBoost to arbitrary norms
- Normalized margin

$$\rho := \min_{1 \le i \le m} y_i f(x_i) = \min_{1 \le i \le m} y_i \sum_{r=1}^{n} \beta_r g_r(x_i)$$

- Nonlinear Programming CLF

$$\max_{\boldsymbol{\beta},\xi,\rho} \quad \rho - \frac{1}{\nu m} \sum_{i=1}^{m} \xi_i$$

$$s.t. \quad y_i \sum_{r=1}^{n} \beta_r f_r(x_i) \ge \rho - \xi_i \quad \forall\ i = 1, ..., m$$

$$\|\boldsymbol{\beta}\|_p^p \le 1, \quad \boldsymbol{\beta} \succeq 0, \boldsymbol{\xi} \succeq 0, \rho \ge 0$$

Multiclass CLF (NLP-*v*MC)

- Novel multiclass CLF
- Margin redefinition

$$\rho_i(x_i,\beta) := \sum_{r=1}^{n} \beta_{(N_C(r-1)+y_i)} g_{r,y_i}(x_i) - \sum_{r=1}^{n} \sum_{j=1, j\neq i}^{N_C} \beta_{(N_C(r-1)+y_j)} g_{r,y_j}(x_i)$$

- Nonlinear Programming (NLP-*v*MC)

$$\max_{\boldsymbol{\beta},\xi,\rho} \quad \rho - \frac{1}{\nu m}\sum_{i=1}^{m}\xi_i$$

$$s.t. \quad \sum_{r=1}^{n} \beta_{(N_C(r-1)+y_i)} g_{r,y_i}(x_i) - \sum_{r=1}^{n} \sum_{j=1, j\neq i}^{N_C} \beta_{(N_C(r-1)+y_j)} g_{r,y_j}(x_i) \geq \rho - \xi_i,$$

$$\|\boldsymbol{\beta}\|_p^p \leq 1, \quad \rho \geq 0, \boldsymbol{\beta} \succeq 0 \quad \boldsymbol{\xi} \succeq 0 \quad \forall i = 1,\dots,m$$

Multiclass CLF (NLP-β)

- Nonlinear Programming-β (NLP-β)

$$\max_{\boldsymbol{\beta},\xi,\rho} \quad \rho - \frac{1}{\nu m}\sum_{i=1}^{m}\xi_i$$

$$s.t. \quad \sum_{r=1}^{n}\beta_r g_{r,y_i}(x_i) - max \sum_{y_j \neq y_i, r=1}^{n} \beta_r g_{r,y_j}(x_i) \geq \rho - \xi_i, \ \forall \ i = 1, ..., m$$

$$\|\boldsymbol{\beta}\|_p^p \leq 1, \quad \beta_r \geq 0, \ \xi_i \geq 0, \rho \geq 0, \quad \forall r = 1, \ldots, n, \forall \ i = 1, \ldots, m.$$

Multiclass CLF (NLP-B)

- Nonlinear Programming-B (NLP-B)

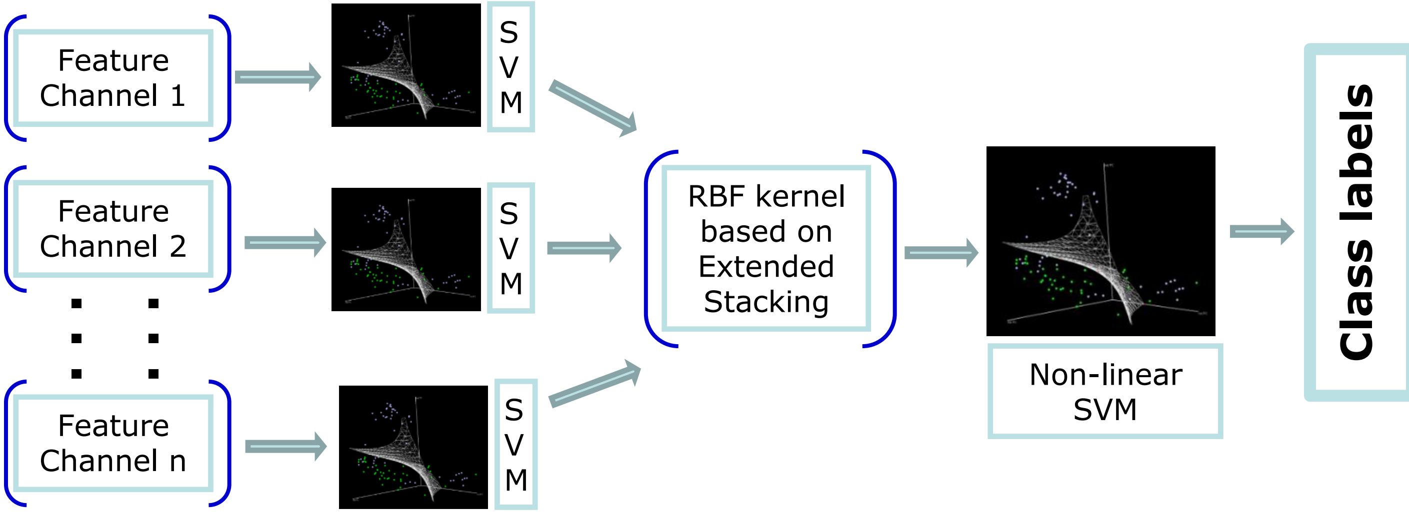
$$\max_{B,\xi,\rho} \quad \rho - \frac{1}{\nu m}\sum_{i=1}^{m}\xi_i$$

$$s.t. \quad \sum_{r=1}^{n} B_r^{y_i} g_{r,y_i}(x_i) - \sum_{y_j \neq y_i, r=1}^{n} B_r^{y_j} g_{m,y_j}(x_i) \geq \rho - \xi_i \;\; i = 1, ..., m,$$

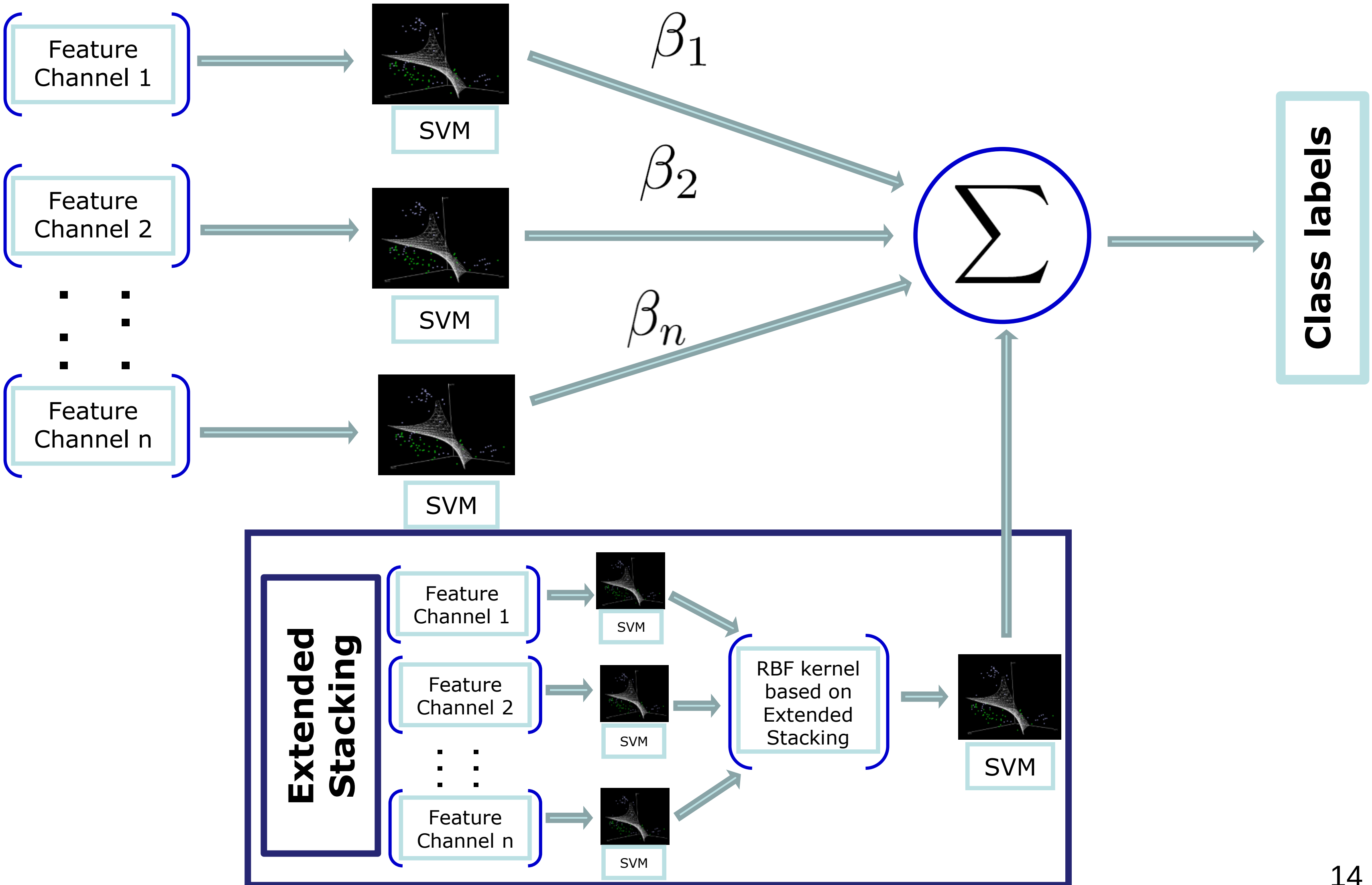
$$\|B\|_p^p \leq 1, \;\; B_r^c \geq 0, \;\; \boldsymbol{\xi} \succeq 0, \rho \geq 0, \;\; \forall \; r = 1, ..., n, c = 1, ..., N_C$$

Extended Stacking

- Break down multiclass problem into 1-vs-all.
  - For each sample its distances from all hyperplanes of 1-vs-all classifier is used as base feature.
- Break down multilabel problem into independent binary problem.
  - For each sample its distances from all hyperplanes of independent binary classifier is used as base feature.



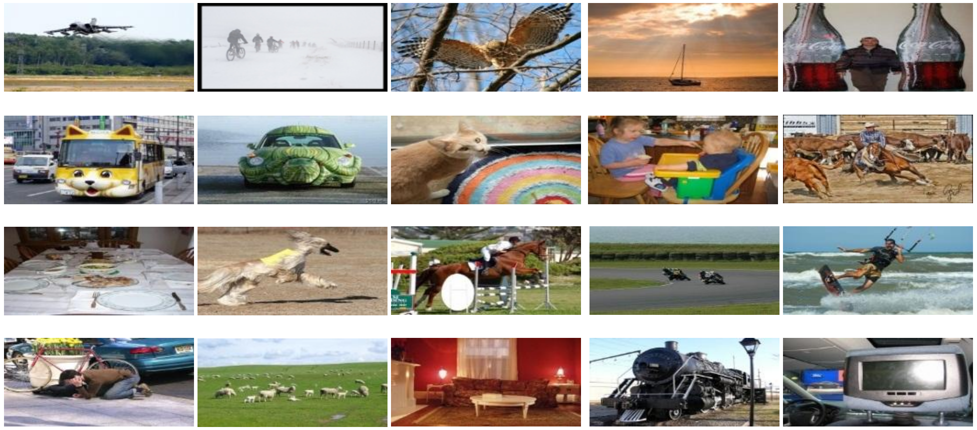
Base plus Stacking Kernel



Results

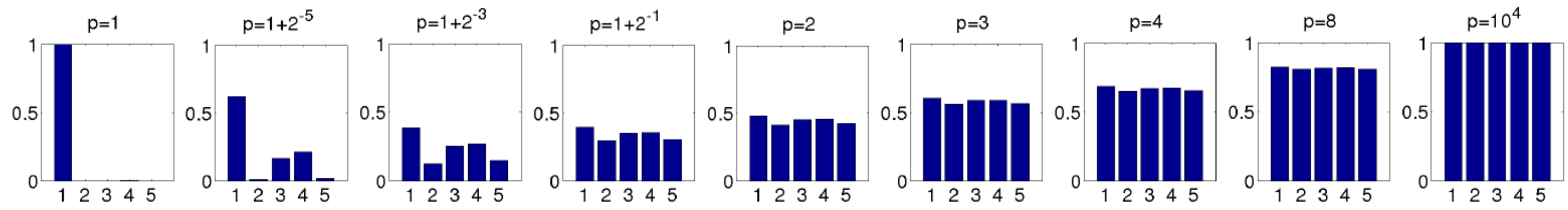
- Multi-Label Dataset (PASCAL VOC2007)
  - 20 classes, 9963 images from internet
  - Mean Average Precision (MAP)

$$MAP = \frac{1}{|R|}\sum_{k=1}^{R} c_k, \qquad c_k = \begin{cases} \frac{|R \cap M_k|}{k} & if \ \ concept \ true \\ 0 & if \ concept \ not \ true \end{cases} \qquad M_k = \{i_1, i_2, \ldots, i_k\}$$



PASCAL VOC 2007

- Feature channel weights learned with various lp-norm for CLF.
  - Sparsity decreases with higher norms



PASCAL VOC 2007

➢ Mean Average Precision of different fusion methods.

Fusion Methods \ norms	1	$1+2^{-3}$	$1+2^{-2}$	$1+2^{-1}$	2	3	4	8	ℓ_∞
MKL	55.42	56.42	58.53	61.07	61.98	62.45	62.61	62.81	62.93
CLF	63.71	63.94	63.97	**63.98**	63.97	63.97	63.77	63.69	63.11
Stacking	**64.44**								
MKL (Base + Stacking)	64.39	64.55	65.06	65.75	66.06	66.23	**66.24**	66.09	65.93
CLF (Base + Stacking)	65.18	65.20	65.45	65.57	65.65	65.63	65.59	65.54	65.48

Results

- Multi-Class Datasets
  - Oxford Flower17 (17 classes)
  - Oxford Flower102 (102 classes)
  - Caltech101 (101 classes)
  - Mean Accuracy
  - Protein Subcellular Localization (4 datasets)
    - 1- MCC in percentage

Mean Accuracy on Oxford Flower17

- Decrease in sparsity led to performance improvement.
- Best results by combining base and stacking kernels.

ML-Methods	1	$1+2^{-3}$	$1+2^{-1}$	2	3	4	8
MKL	87.2±2.7	74.9±1.7	72.2±3.6	71.2±2.7	70.6±3.8	73.1±3.9	81.0±4.0
NLP-β	86.5±3.3	86.6±3.4	86.6±1.1	86.7±1.2	87.4±1.5	**87.9±1.8**	87.8±2.1
NLP-νMC	85.5±1.3	86.6±2.0	87.6±2.2	87.7±2.6	**87.8±2.1**	87.7±2.0	87.8±1.9
NLP-B	84.6±2.5	84.6±2.4	84.8±2.6	84.8±2.5	85.5±3.7	86.9±2.7	87.3±2.7
Stacking	**89.4 ± 0.5**						
MKL(Base +Stacking)	89.3±0.9	79.7±2.7	77.6±1.2	74.7±2.4	73.8±2.6	77.8±4.3	86.3±1.9
NLP-β(Base +Stacking)	**90.2±1.5**	89.3±0.7	89.6±0.5	89.2±1.6	89.3±1.2	89.1±1.4	89.0±1.0
NLP-νMC(Base +Stacking)	86.1±2.5	87.3±1.4	88.5±0.5	88.6±0.9	88.6±0.9	88.8±1.1	88.9±1.2

Comparison with State-of-the-Art	
MKL-FDA (ℓ_p) [Yan et al. CVPR10](best state-of-the-art using 7 kernels)	86.7±1.2
MKL-avg (ℓ_∞)(using 7 kernels)	84.9±1.9
CLF (ℓ_∞) (using 7 kernels)	86.7±2.7

Summary of Mean Accuracy on Computer Vision Datasets

- Similar trend was observed for oxford flower102 and clatech101.
  - Summary of results is given in table (more details in paper).

ML-Methods	Oxford flower 17	Oxford flower 102	caltech101
state-of-the-art-best	86.7±1.2	72.8	68.6±2.2
proposed Extended Stacking	89.4±0.5	77.7	68.0±2.4
proposed-CLF-best	**90.2±1.5**	**80.3**	**70.7±1.9**

Protein Subcellular Localization

- Results on 4 bioinformatics datasets validate our experiments.
  - Prediction error is measured as 1- Matthew Correlation Coefficient (MCC) in percentage.

ML-Methods	plant	nonpl	psortNeg	psortPos
SVM-best [kloft et al. JMLR11]	8.18±0.47	8.97±0.26	9.87±0.34	13.01±0.63
FDA-best [Yan et al. JMLR11]	10.85±2.37	10.84±1.72	9.74±2.00	12.59±4.10
proposed-best	**5.25±1.88**	**6.39±1.12**	**9.16±1.67**	**10.40±3.56**

Conclusions

- A novel nonlinear separable convex optimization formulation for multiclass classifier fusion.
  - Learns weight of each class in each feature channel.
- Arbitrary norm in the existing CLF formulation
  - Don't reject informative channels
  - Robust against noisy and redundant channels
  - Norm can be learnt using validation set
- Extended stacking for binary and multiclass problems.
  - Stacking plus base kernels gives best results
- Extensive evaluation on challenging datasets.

Thanks!

Questions ???