



# KERNEL FUSION FOR VIDEO RETRIEVAL TASKS

Pascal Michailat and Slav Petrov  
University of California at Berkeley

## Context of the TRECVID Challenge

Video annotation is an easy task for humans, but researchers have been struggling for decades with the problems of:

- Automatic speech recognition, text retrieval, object recognition.

The NIST has been conducting an annual Text REtrieval Conference since 1992, and additional VIDEO tracks were added in 2001: **TRECVID challenge**.

- Encourage automatic segmentation, indexing, and content-based retrieval from large video corpora.

## Introduction

Presentation of a comprehensive statistical framework for information retrieval on video shots.

- Design of language and video features.
- Binary classification in 39 categories with 1-norm soft-margin SVM.
- Exploration of the SVM hyper-parameter space:
  - Regularization parameter: trade-off margin error vs. margin width.
  - Cost asymmetry: trade-off recall (false negative) vs. precision (false positive).
- Data fusion using a multiple kernel SVM algorithm.

Data set constituted of video shots from TV broadcast news manually annotated.

- 99 hours of news in English, 32 hours of news in Chinese, and 83 hours of news in Arabic.
- Implementation of our algorithm on a small scale problem:
  - Categories with at least 15 words per shot on average: Chart, Computer-TV-Screen, Corporate Leader, Flag US, Government Leader, Map, Studio, Weather.
  - Select 100 positive examples for each category for training and for testing:  $\approx 500$  training and testing examples.

## Vision Feature Space

Illustration of GB features and correspondence between two images (courtesy of A. Berg):

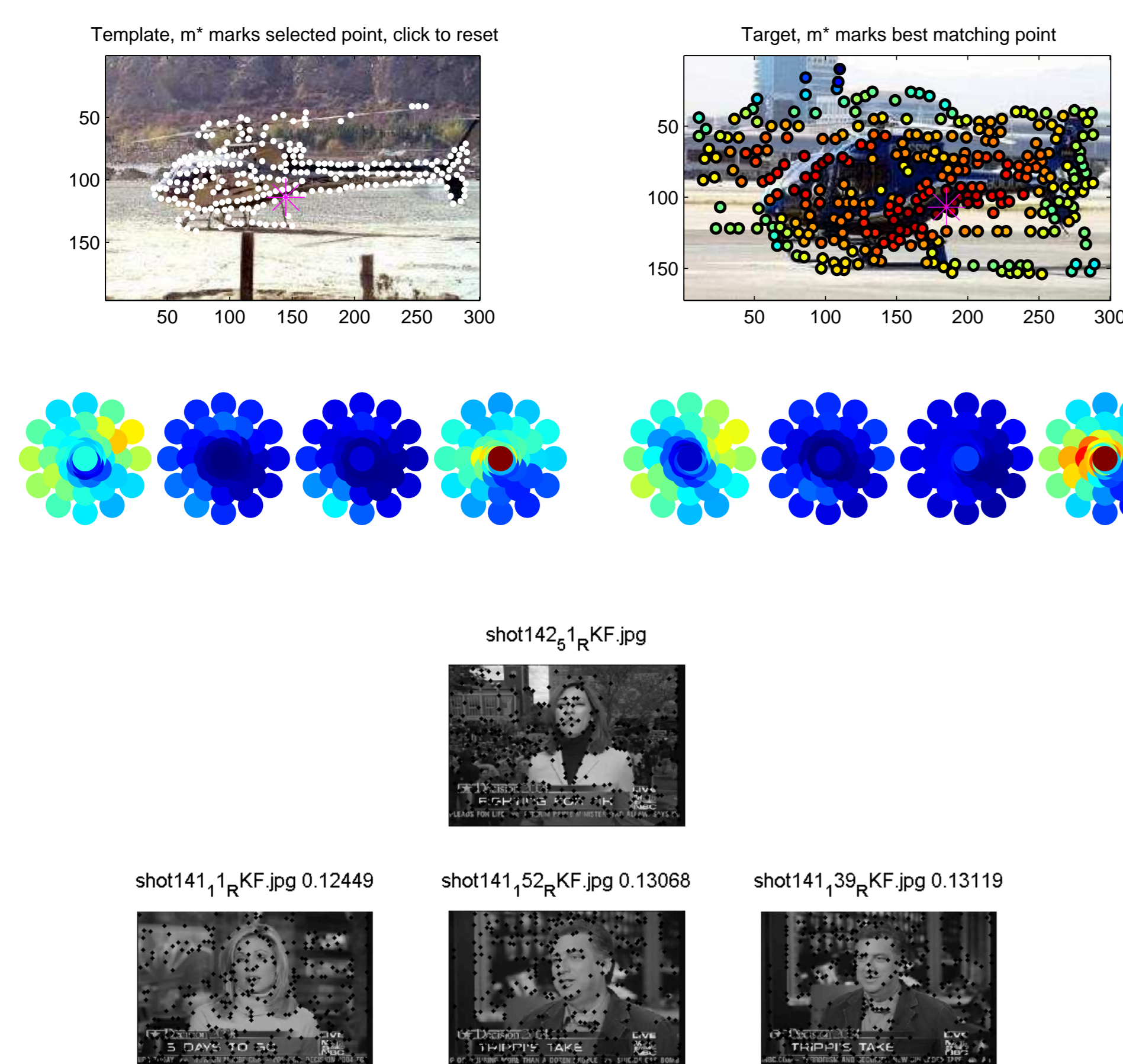


FIGURE 1: Query image at the top and closest matches in the bottom row. The centers of the GB features are marked. Unfortunately, best matches are created by the captions.

Regular 1-norm soft-margin formulation:

$$\min_{w,b,\zeta} \frac{1}{2}w^T w + \frac{c}{n} \sum_{i=1}^n \zeta_i$$

$$\text{subject to } y_i (w^T x_i + b) \geq 1 - \zeta_i \quad i = 1 \dots n$$

$$\zeta_i \geq 0 \quad i = 1 \dots n$$

Hinge loss function:  $l(y, \hat{y}) = (1 - y\hat{y})_+$ . Activation function:  $g(x_i) = w^T x_i + b$ .

<b>Regularization</b>	<b>Regularization and cost asymmetry</b>
$\hat{L}(w, b) = \frac{c}{n} \sum_{i=1}^n l(y_i, g(x_i))$	$\hat{L}(w, b) = \frac{c}{n} (\gamma \sum_{i y_i=+1} l(y_i, g(x_i)) + (1 - \gamma) \sum_{i y_i=-1} l(y_i, g(x_i)))$

## Multiple Kernel Learning with Soft-Margin SVM

Kernel matrix of soft-margin SVMs chosen from a family of  $m$  kernel matrices  $(K_1, \dots, K_m)$  in a semi definite programming setting. Optimizes both the kernel matrix and the decision boundary  $(w, b)$  simultaneously: convex, non-smooth optimization over a convex set.

<b>Constraint on kernel</b>	<b>Multiple kernel SVM dual formulation</b>
$K = \sum_{i=1}^n \mu_i K_i$	$\min_{\alpha_i \in \mathbb{R}} \sum_{i=1}^n \alpha_i + \zeta$
$\mu_i \geq 0 \quad i = 1, \dots, n$	subject to $0 \leq \alpha_i \leq C \quad i = 1 \dots n$
$c = \text{tr}(K)$	$\alpha^T y = 0$
	$\alpha^T D(y) K_j D(y) \alpha \leq \frac{\text{tr}(K_j)}{c} \zeta \quad j = 1 \dots m \quad (1)$

$\mu_j$  recovered as Lagrange multipliers for the last set of  $m$  constraints (1).

1. **Kernel Tuning:** One source of information available. Which kernel shape: linear, polynomial, radial basis function? what values of the parameters?  $\Rightarrow$  Find the optimal linear combination of these kernels and eliminate “useless” kernels.
2. **Data Fusion:** Several heterogeneous and complementary sources of information available for the classification task  $\Rightarrow$  Optimal data fusion in one step without meta-classifier.

## Language Feature Space

Generate language features along two dimensions:

- Varying the word count: term frequency (TF), log term frequency (LTF), inverse document frequency (IDF). We try to reduce the importance of very common words.
- Transforming the initial features: normalization (N), suppression of the most common words (R), both (NR). We try to homogenize the variance along every dimension.

Weight  $\omega$  associated by the SVM to a given word  $m$  in document  $d$ :

$$\omega(m) = \sum_{d \in D} \alpha_d y_d \text{Count}_d(m)$$

The top ten words learned by SVM as strong indicators for and against the class Weather:

No.	Pos. Word	Weight	Neg. Word	Weight
1.	rain	29979	my	-12814
2.	up	27033	president	-12592
3.	moisture	19580	percent	-11935
4.	way	19126	four	-11871
5.	coming	18922	election	-11551
6.	coast	18469	democrats	-10678
7.	west	17892	can	-10234
8.	temperatures	17585	state	-9841
9.	showers	17218	hundred	-9834
10.	winds	17093	for	-9783

## Combining Language and Vision

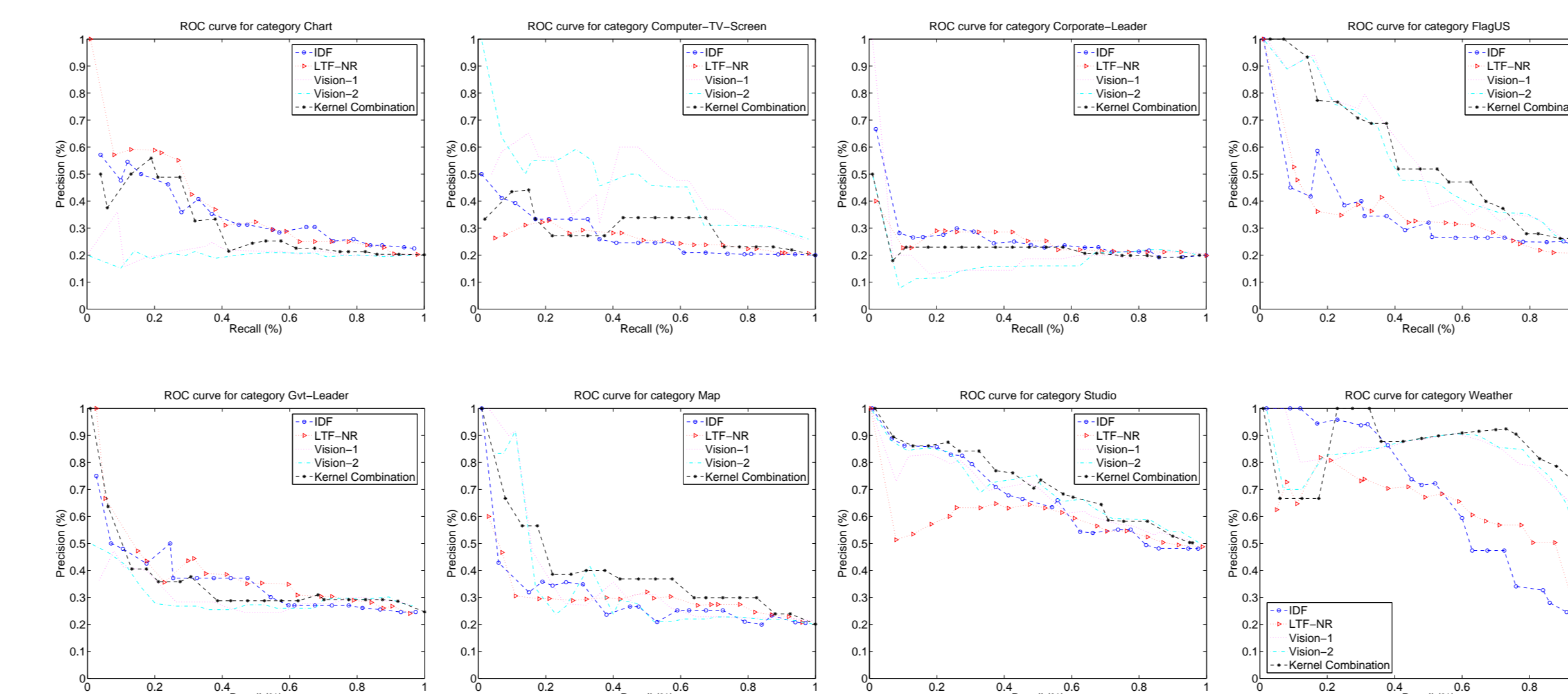
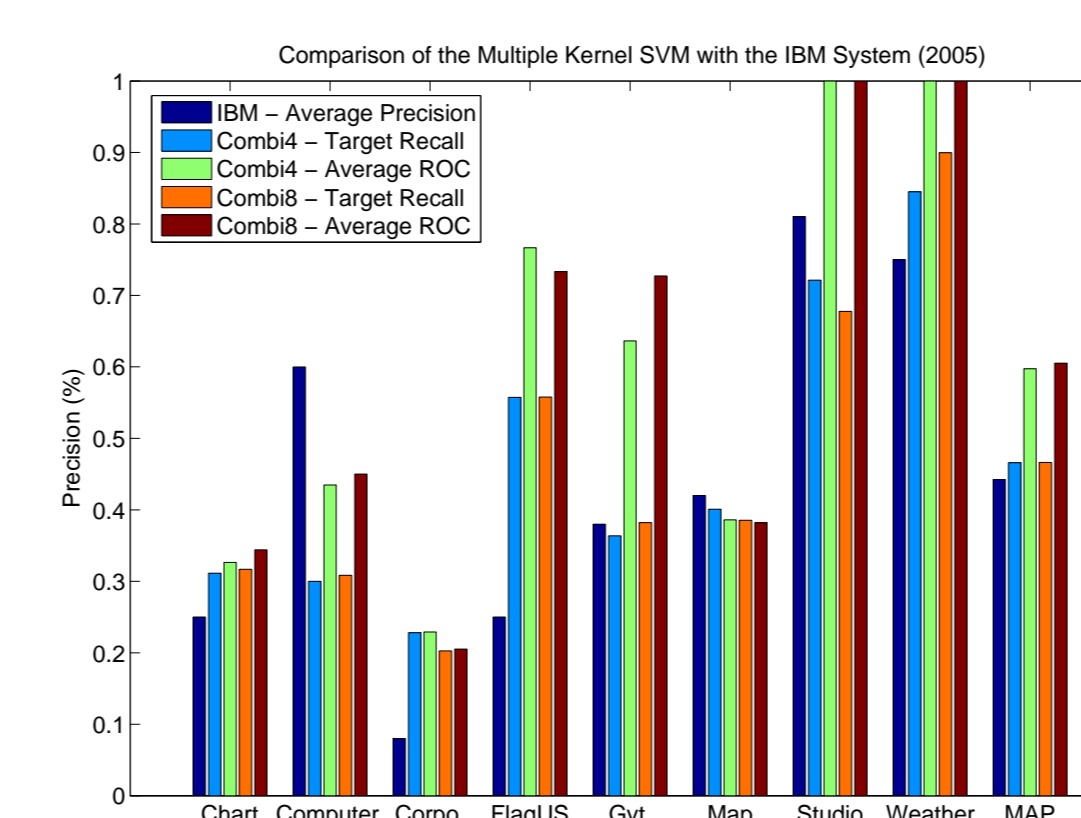


FIGURE 2: Performance of the fusion of 2 language and 2 vision kernels on 8 categories.

**IBM:** Precision score = AP on the first 1,000 examples classified in each category.

**Kernel Fusion:**

- Combi4: combinaison of Vision-1, Vision-2, IDF and LTF-NR.
- Combi8: all of the preceding features and IDF-R, IDF-NR, IDF-N, and TF.
- Upper bound: precision on testing set with projected target recall for TRECVID.
- Lower bound: average of the ROC curve.



## Scaling Up the Multiple Kernel SVM

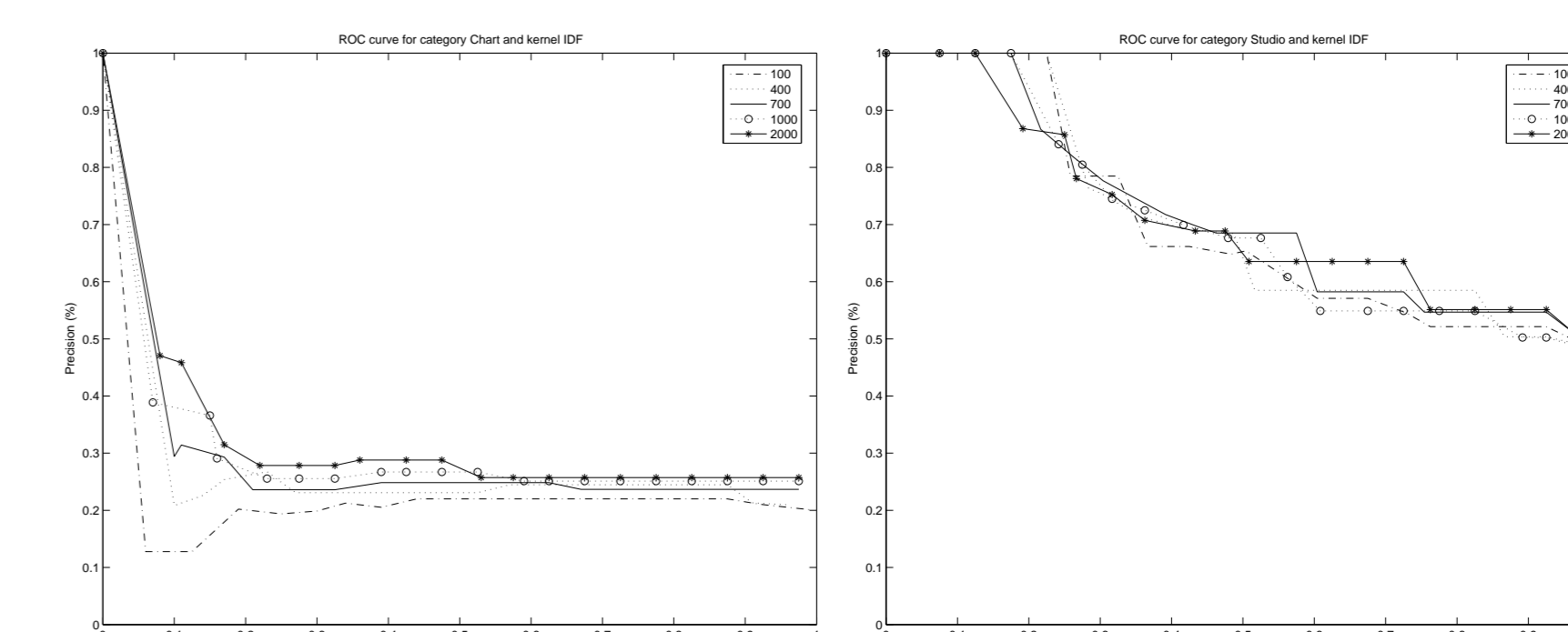


FIGURE 3: Impact of the number of negative examples in the training set on the performance of the SVM on a fixed test set. The improvements are limited above 700 negatives.

## Next Steps



**ASR from TRECVID:** .. state trooper injured in a car crash .. jefferson was seen the eastbound lanes of the inter - upper moral intention of .. the officer was taken to an internal morale hospital that the centers are not life threatening were told .. to lanes new-line but remain closed while police are investigating ..

**ASR from SRI-ICSI:** tonight a pennsylvania state trooper injured in a car crash chopper ten over the scene the eastbound lanes of the p. h. n. pike in upper moreland township the officer was taken to abington memorial hospital the troopers injuries are not life threatening we're told two lanes of the turnpike remain closed while police are investigating