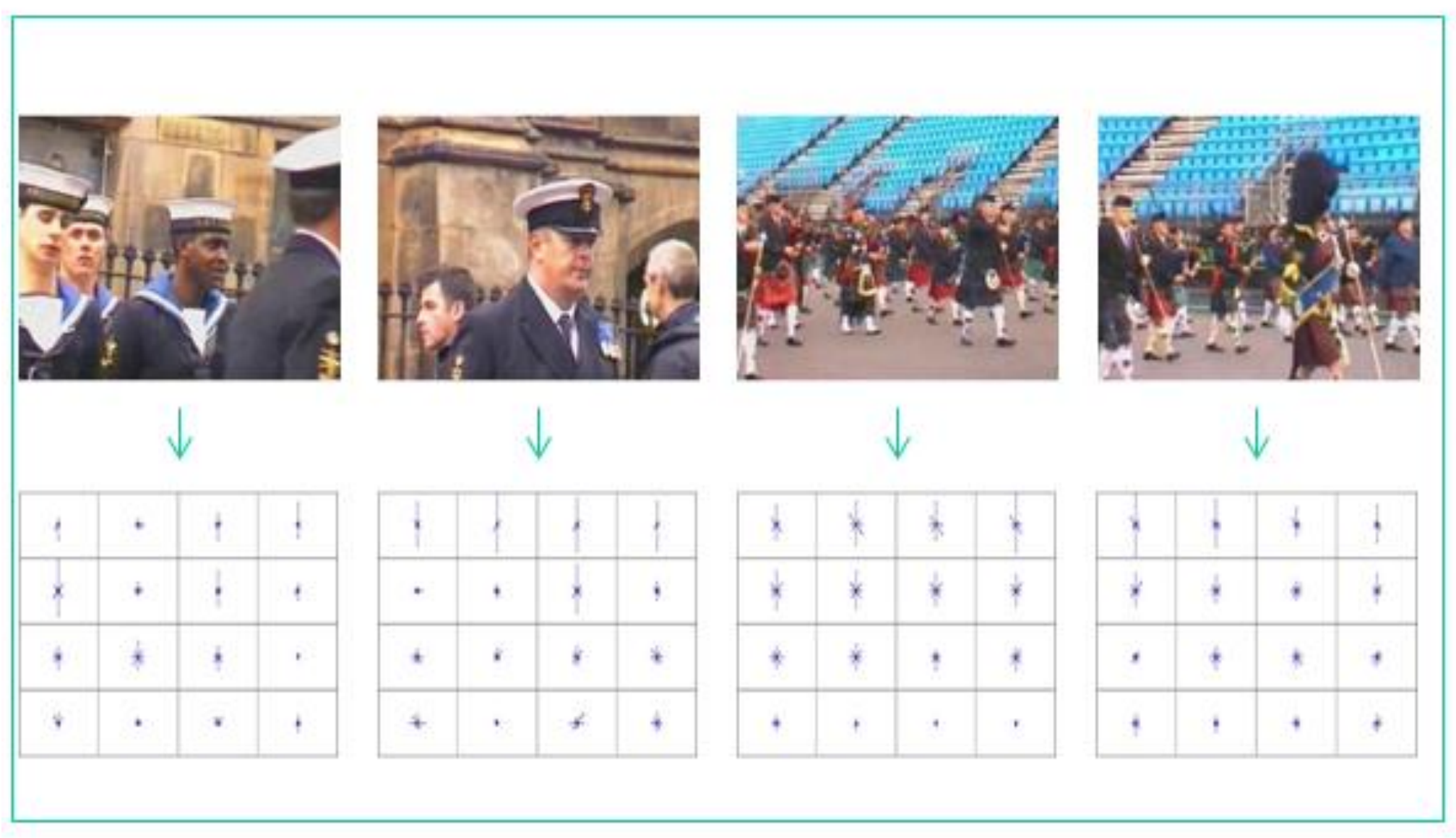
# Time Matters!
# Capturing Variation in Time in Video using Fisher Kernels

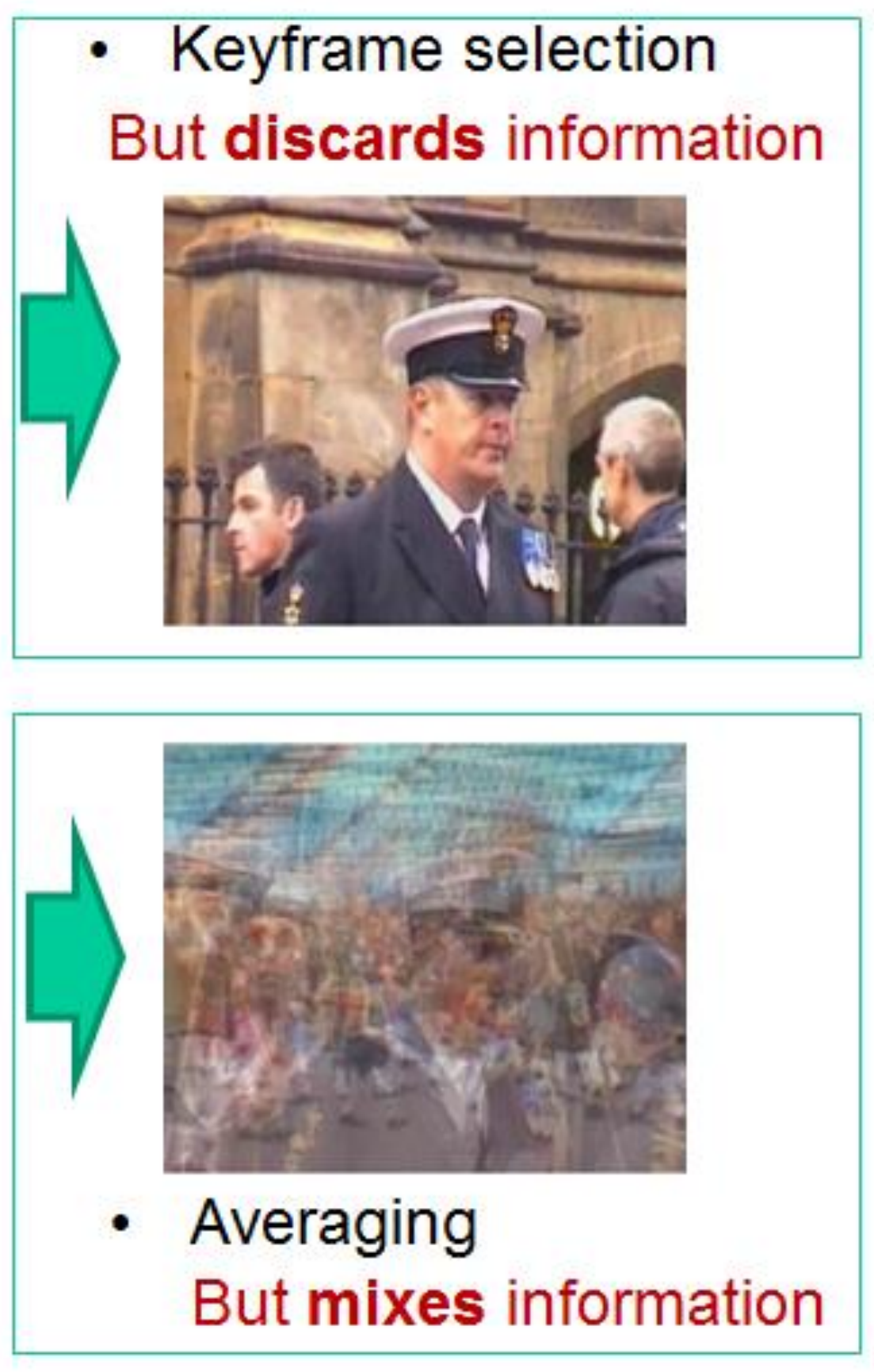
***Ionuț Mironică,[1] Jasper Uijlings,[2] Negar Rostamzadeh,[2] Bogdan Ionescu,[1,2] Nicu Sebe[2]***

[1] **LAPI – University „POLITEHNICA” of Bucharest, 061071, Romania**
[2] **MHUG DISI – University of Trento, Italy**

## Problem: Aggregating features over time

**Traditional methods to aggregate frame-based features**





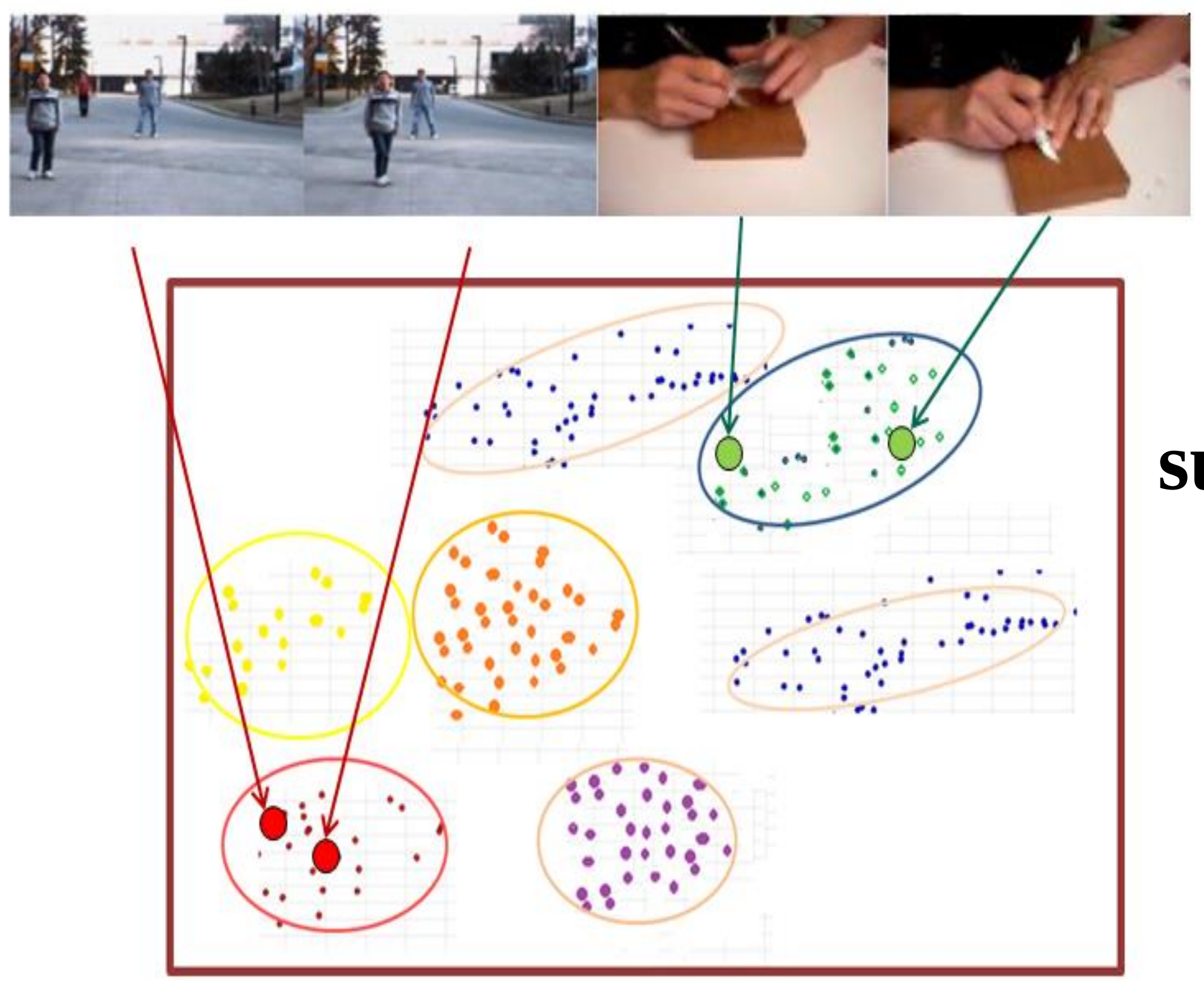
- Keyframe selection

But **discards** information

- Averaging

But **mixes** information

## Solution: Fisher Kernels for Variation in Time



**Temporal order is lost**

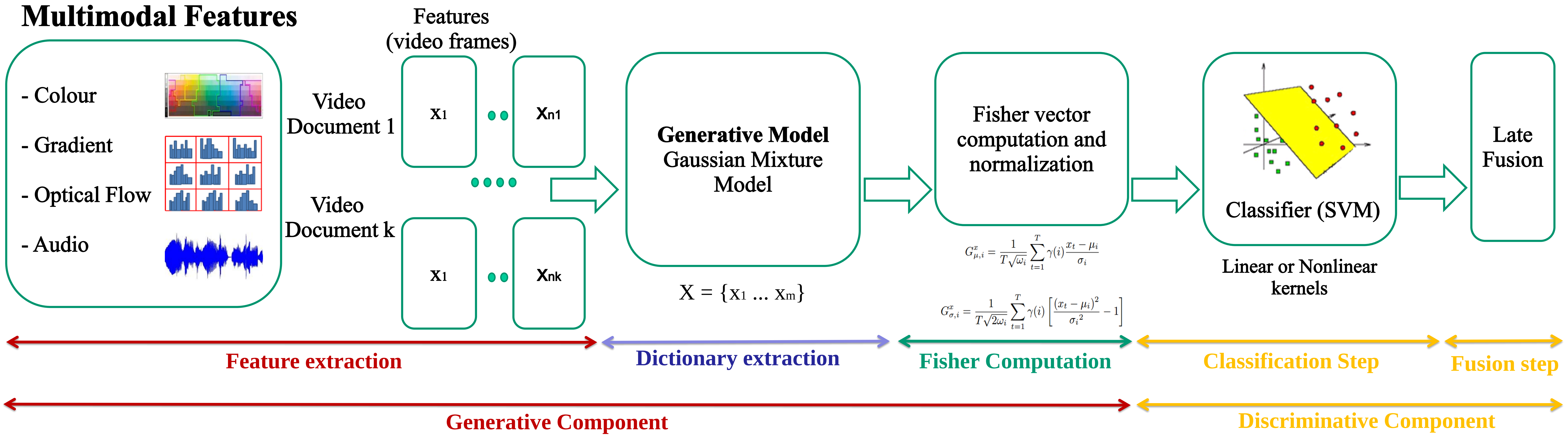
**but captures**

**subtle variations in time within a cluster**

**drastic variations in time by different clusters**

## Fisher Kernel Framework for Variation in Time

- **combines the benefits of generative and discriminative approaches**
- **represents a signal as the gradient of the probability density function that is a learned generative model of that signal**



**Multimodal Features**

- Colour
- Gradient
- Optical Flow
- Audio

Features (video frames)

Video Document 1: x1 •• xn1

••••

Video Document k: x1 •• xnk

**Generative Model** Gaussian Mixture Model

X = {x1 ... xm}

Fisher vector computation and normalization

$$G^{x}_{\mu,i} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma(i) \frac{x_t - \mu_i}{\sigma_i}$$

$$G^{x}_{\sigma,i} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \gamma(i) \left[ \frac{(x_t - \mu_i)^2}{{\sigma_i}^2} - 1 \right]$$

Classifier (SVM)

Linear or Nonlinear kernels

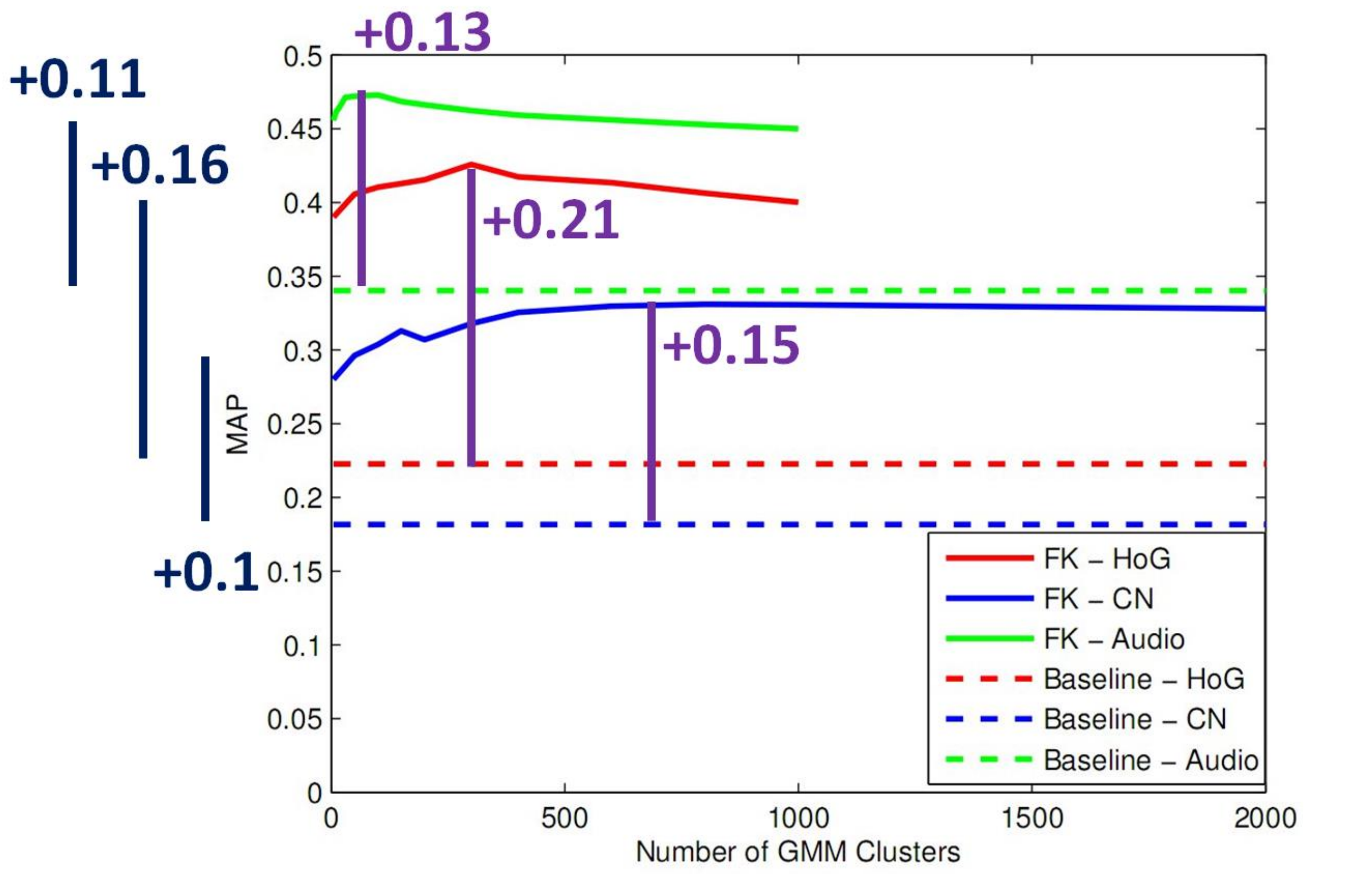
Late Fusion

**Feature extraction** | **Dictionary extraction** | **Fisher Computation** | **Classification Step** | **Fusion step**

**Generative Component** | **Discriminative Component**

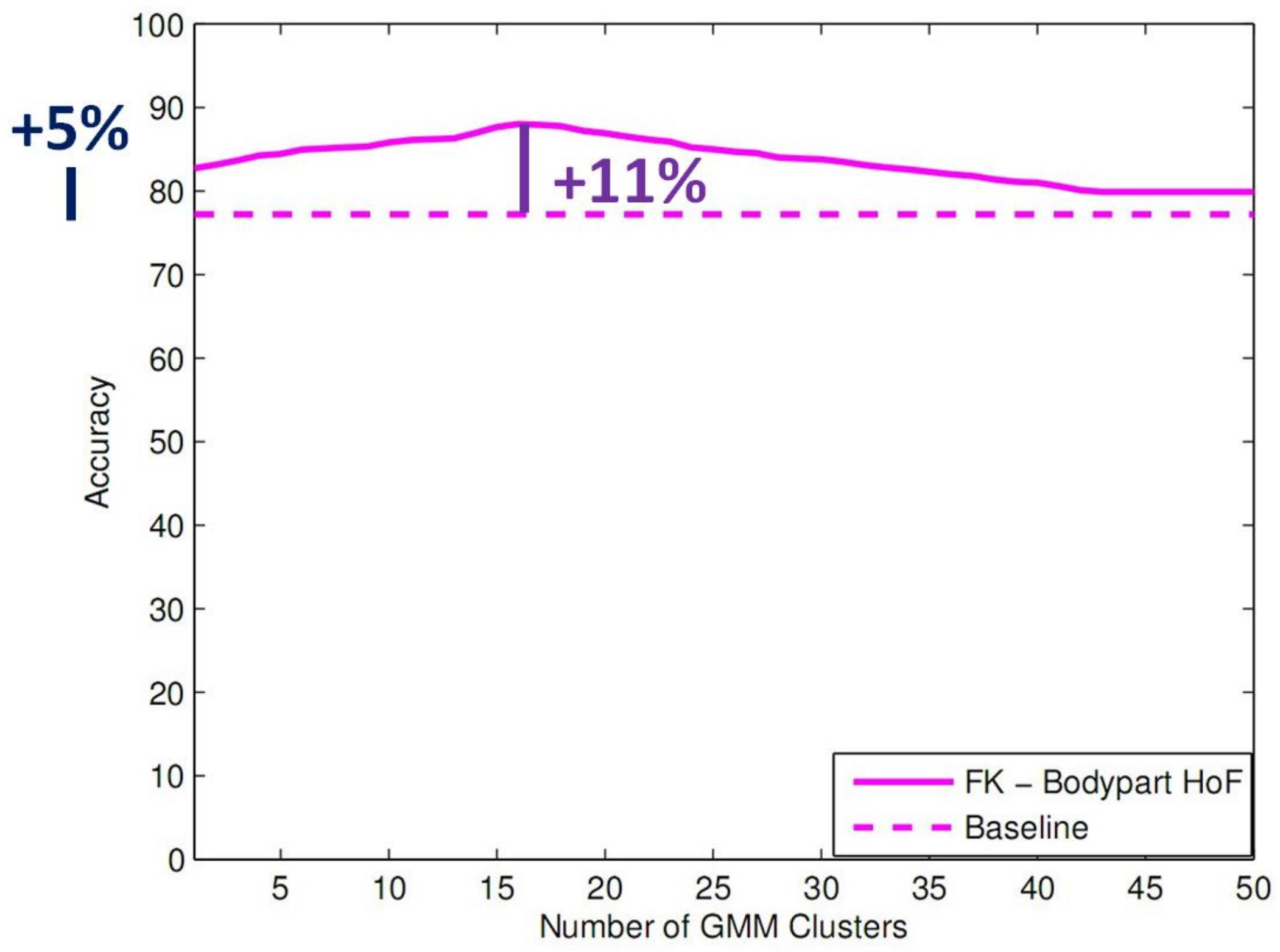
## Experimental Results

### MediaEval 2012 dataset- Genre Retrieval



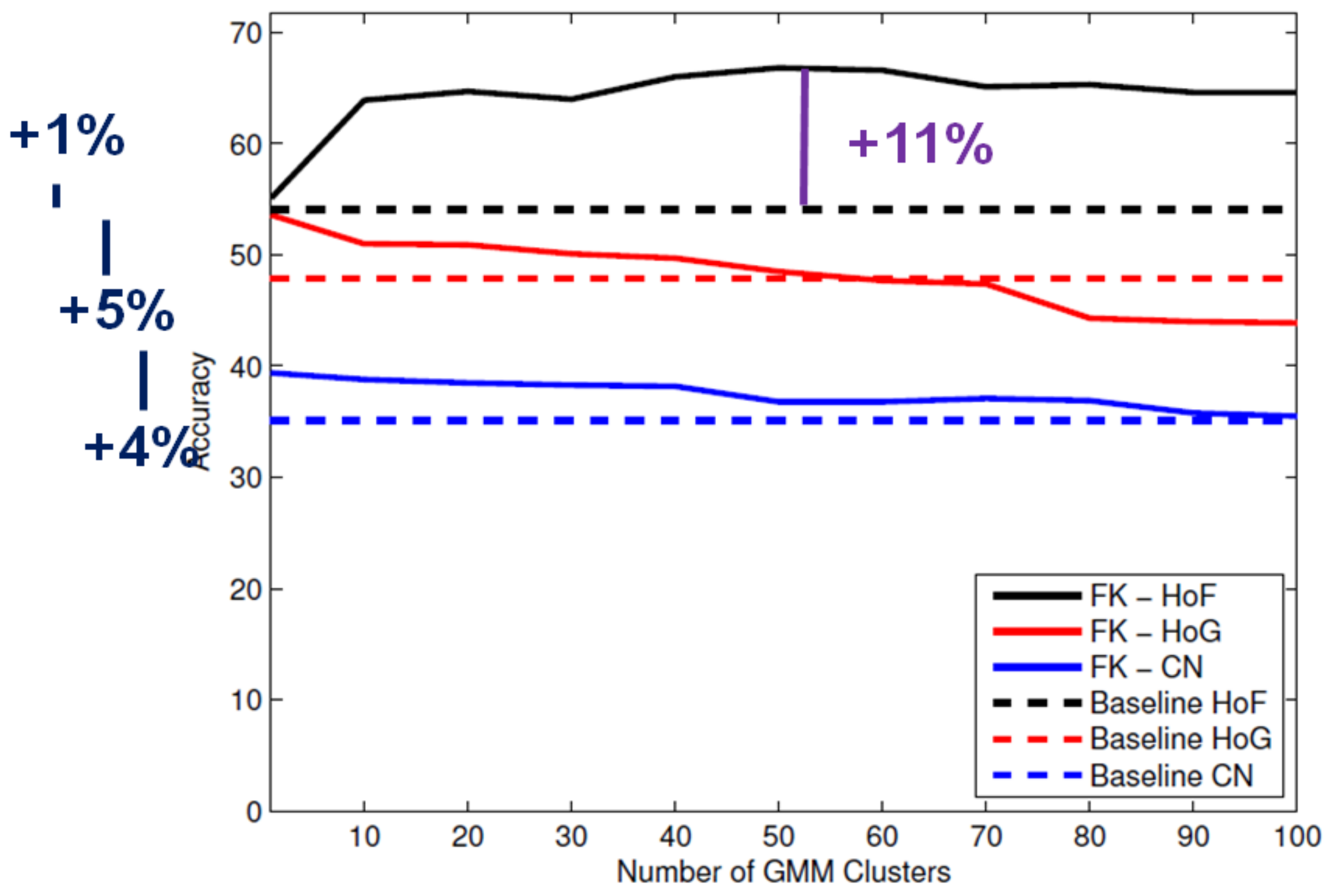
| Feature type | Summary SoA method MediaEval 2012 | MAP SoA | MAP ours | |
|---|---|---|---|---|
| Audio | Block Based Audio Features and 5-NN [6] | 0.192 | 0.475 | **+0.28** |
| Visual | Visual descriptors (Color, Texture, rgbSIFT) [23] | 0.350 | 0.460 | **+0.11** |
| Audio & Visual | - | - | 0.550 | |
| Metadata & Text ASR | BoW Text ASR & metadata [20] | 0.523 | - | |

### ADL dataset - Daily Activities



| Method | Accuracy | |
|---|---|---|
| **This paper** | **97.3%** | **+1.3%** |
| Wang et al. [30] | 96.0% | |
| Lin et al. [11] | 95.0% | |
| Messing et al. [14] | 89.0% | |

### UCF dataset – Sport Recognition



| Method | Accuracy |
|---|---|
| Reddy et al. [17] | 76.9% |
| **This paper** | **74.7%** |
| Solmaz et al. [26] | 73.7% |
| Everts et al. [5] | 72.9% |
| Kliper-Gross et al. [8] | 72.6% |
| Solmaz et al. [26]: GIST3D | 65.3% |

**Large improvements** using **single cluster** only **& Larger improvements** using **multiple clusters**

**Conclusion: State-of-the-art** results or **better** using **cheaper** features!