Eye movements on natural videos: Predictive power of different low-level features

Eleonora Vig, Michael Dorr, Thomas Martinetz, and Erhardt Barth
Institute for Neuro- and Bioinformatics, University of Lübeck, Germany
{vig|dorr|martinetz|barth}@inb.uni-luebeck.de, http://www.inb.uni-luebeck.de

Motivation

Eye movements in dynamic scenes are influenced by low-level image properties such as contrast, edge density, motion, or color. We computed these features on a spatio-temporal multiresolution pyramid and analyzed their predictability by training Machine Learning algorithms with them. The error rates indicate what impact a certain low-level feature has on guiding eye movements. The current research contributes to the better understanding of the nature of salient events. The long-term objective of the GazeCom project is to improve visual communication by guiding the gaze with real-time interactive and gaze-specific display of information.

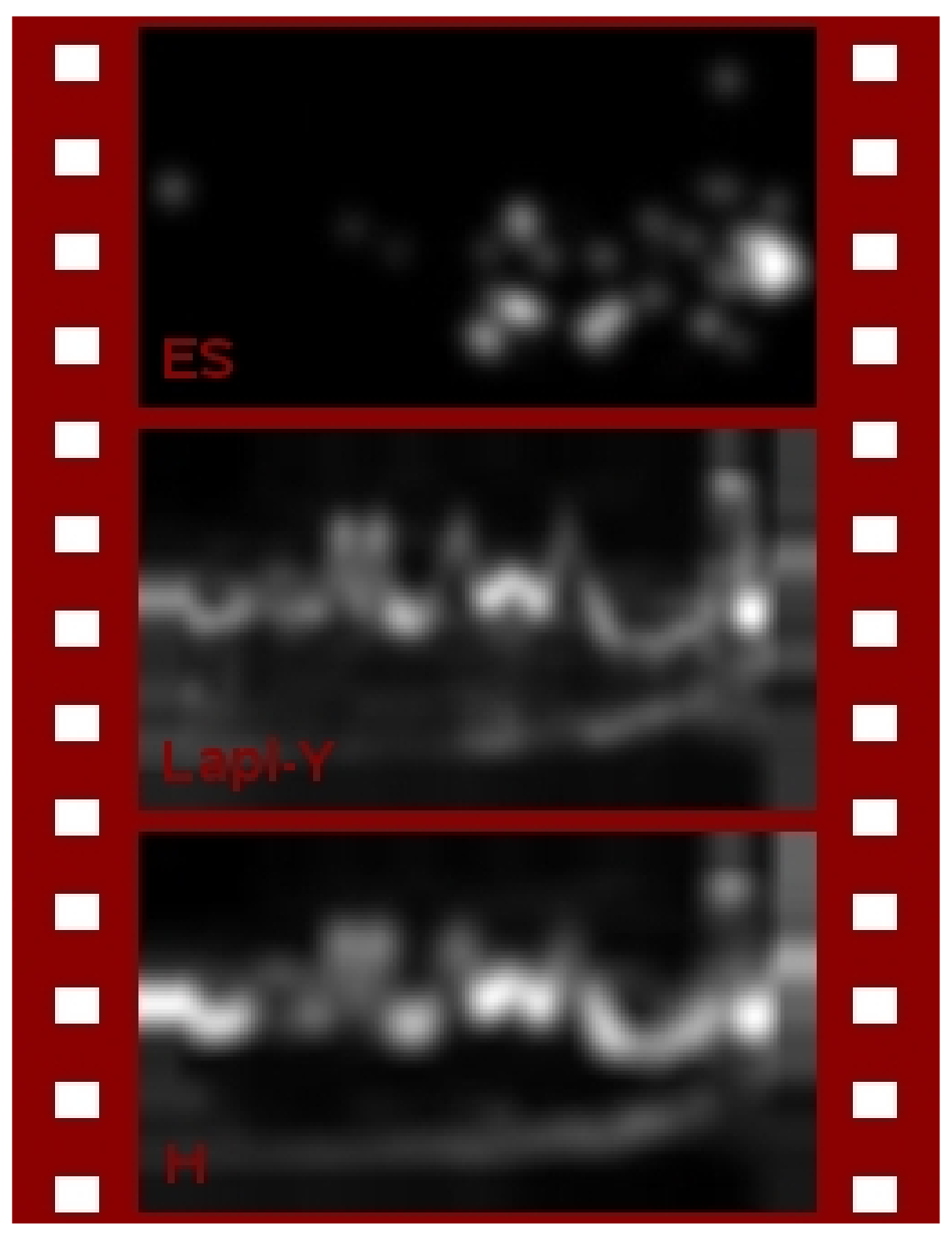
Data & Representations

Data set

A large data set of gaze samples was obtained from 54 subjects watching 18 short, high-resolution videos (20 s duration each) showing outdoor scenes. For obtaining a data set of movie-blocks labeled as “salient” and “non-salient”, an empirical saliency measure was defined as the density of saccade landing points, by placing a 3-D Gaussian at each gaze sample of the observers. The superposition of these Gaussians resulted in the saliency map. The salient (non-salient) block centers were then obtained by picking a value higher (lower) than a predefined threshold on this map. From each of the 18 video clips we cut local movie blocks of 17 x 17 pixels and 8 frames, 2000 clips per class.

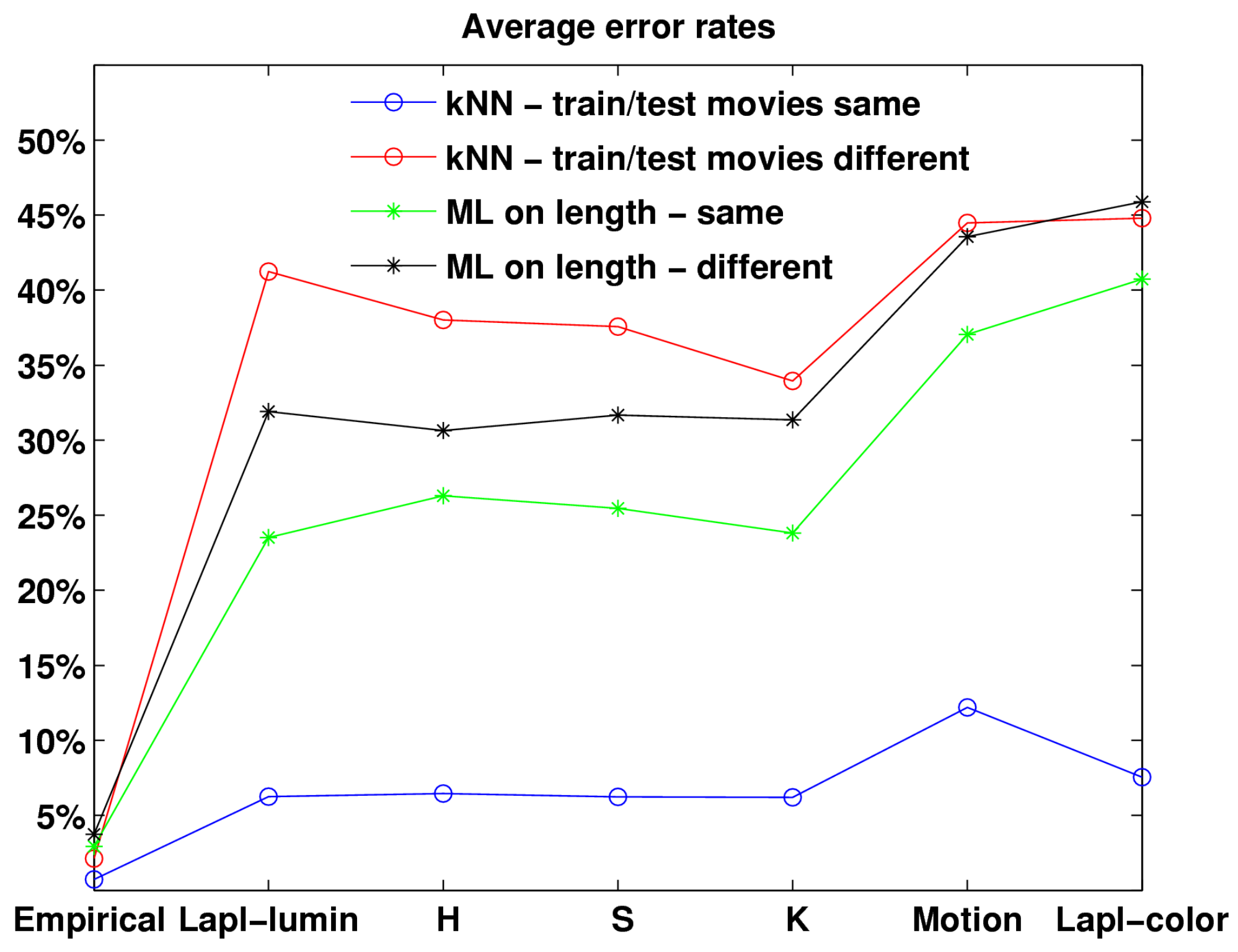
Representations

- **Empirical saliency (ES)**: as described in "Data set".
- **Laplacian**: low-pass filtered images of the third level of a Laplacian pyramid (1.7-3.4 cycles/degree frequency range) built on the Y (luminance) color plane of the images.
- **Color opponency**: same as above, only on U (chrominance) color plane of the images (red/green opponency).
- **Motion**: estimated on the third level of a spatio-temporal pyramid using the structure tensor J.
- **Geometrical invariants H, S, K**: analytical saliency measures, indicating the intrinsic dimensionality of the signal.

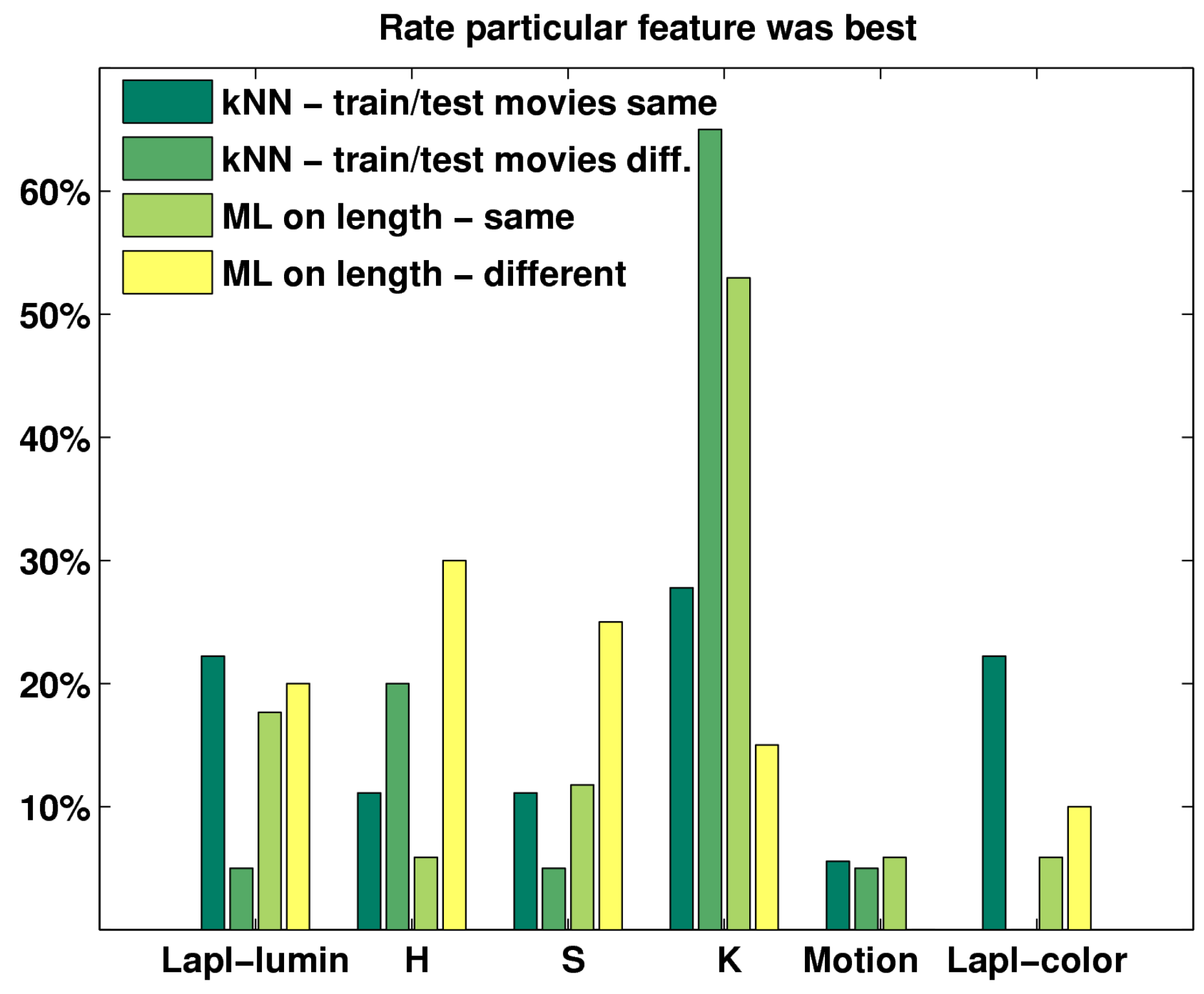


Two different classifiers (maximum likelihood on feature-vector length and k-nearest-neighbor on full feature vectors) were used to classify the movie blocks into the two classes ("salient" and "non-salient") for all representations.

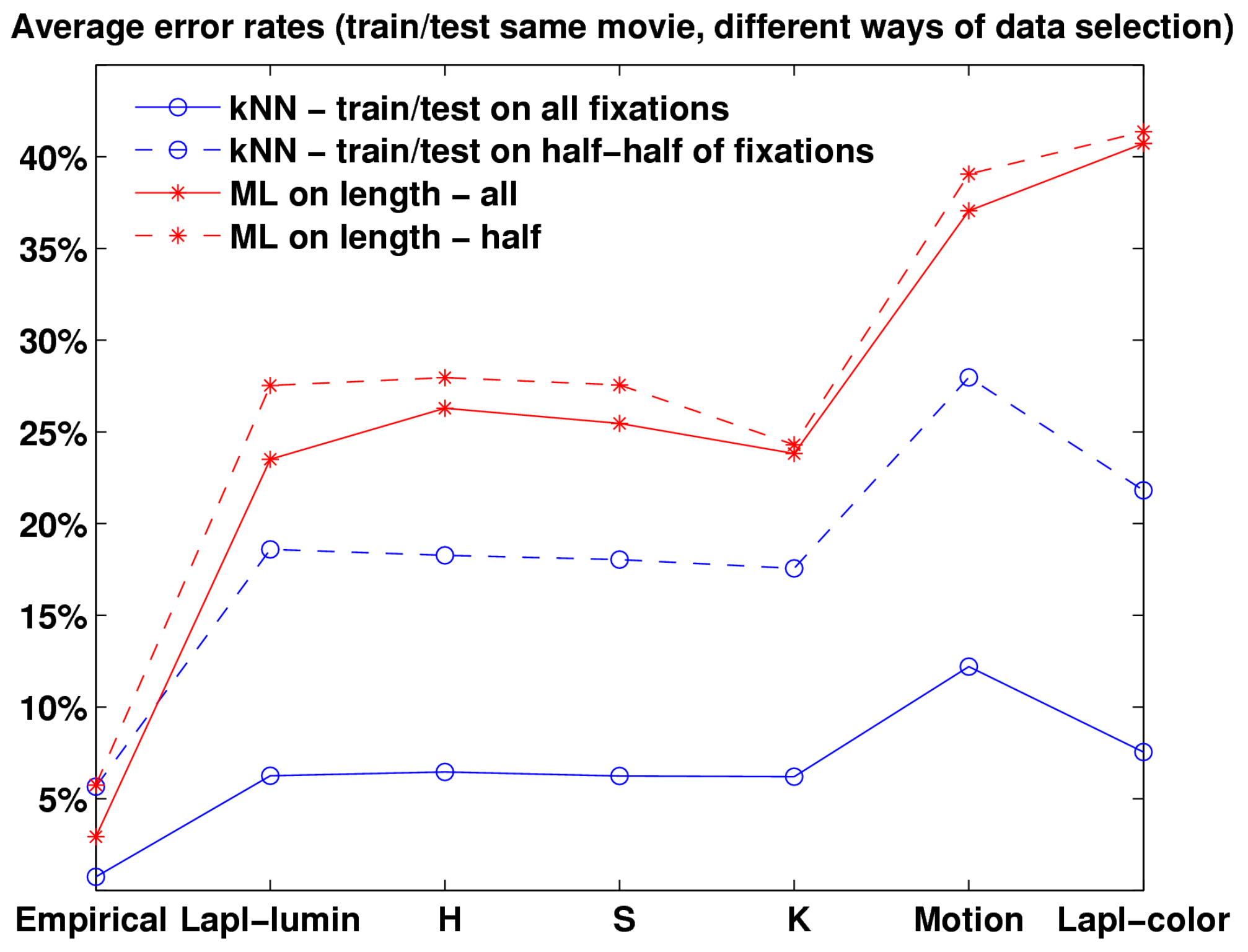
Results



Averages of classification error rates for different representations. First, the two classifiers were trained and tested on the same movie, then trained on several movies and tested on a different one.



Rate of movies when a particular feature gave the lowest classification error. E.g., K was the best feature to use in more than 60% of the movies when kNN was the classifier, tested and trained on different movies.



Average error rates when training and testing were performed on the same movie and (1) all subjects were used for training/test (2) training data was taken from half of the subjects, test data from the other half.

Conclusions

- The low error rate (6%) obtained for a kNN trained/tested on the same movie doesn't generalize to the case when a set of different training movies is used. In the ML case there is only a small increase in error rate.
- K performs slightly better than other representations, however the difference is not significant.
- kNN error rate increases by a factor of 3 generalizing across subjects; only slight impact in ML case.
- Error rates are highly movie dependant.

Our research has received funding from the European Commission within the GazeCom project (IST-C-033816) of the FP6.
All views herein are those of the authors alone; the European Community is not liable for any use made of the information.