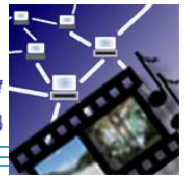


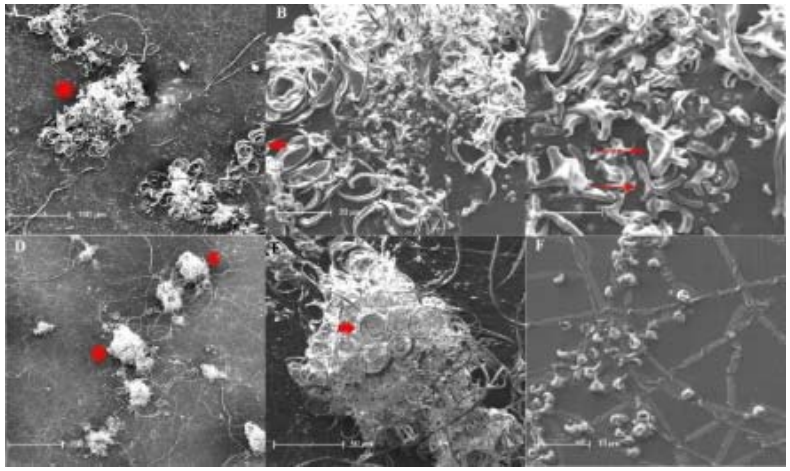
Compound Figure Separation Combining Edge and Band Separator Detection

Mario Taschwer, AAU, Klagenfurt, Austria
Oge Marques, FAU, Boca Raton, FL, USA

FAU
FLORIDA ATLANTIC UNIVERSITY

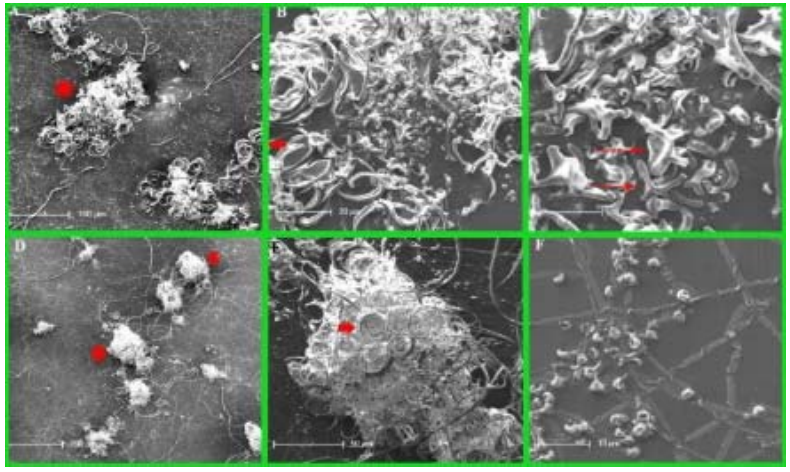
January 5, 2016
MMM 2016

PROBLEM

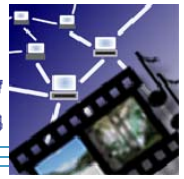


Compound image in scientific article

Compound figure separation (CFS) – automatic!



segmented into subfigures

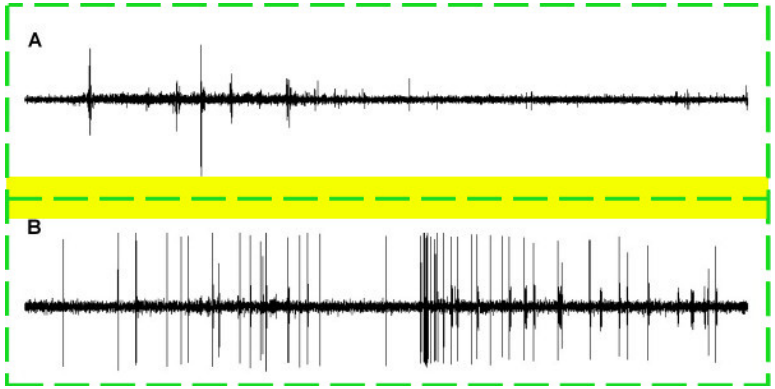


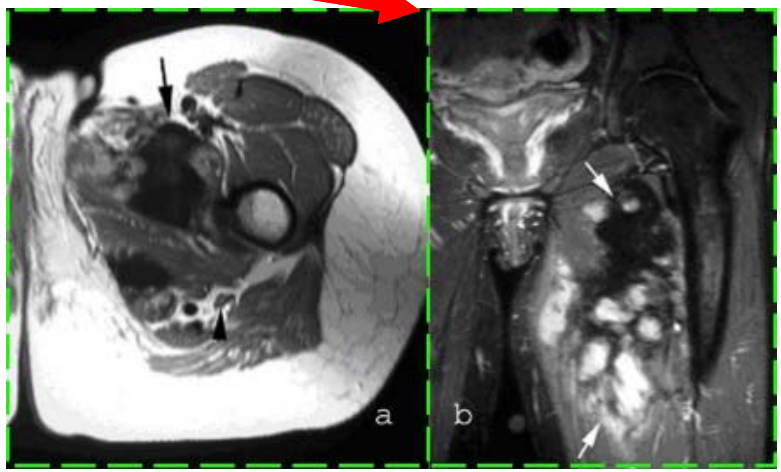
MOTIVATION

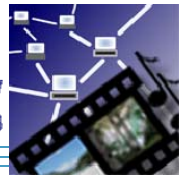
- Biomedical literature:
  - 40%-60% of figures in articles are compound
  - infeasible to separate subfigures manually
- Compound images hinder
  - Content-based analysis
  - Content-based indexing for retrieval
- CFS recognized as research problem recently
  - research fostered by ImageCLEF CFS tasks in 2013 and 2015 (biomedical domain)

KNOWN APPROACHES

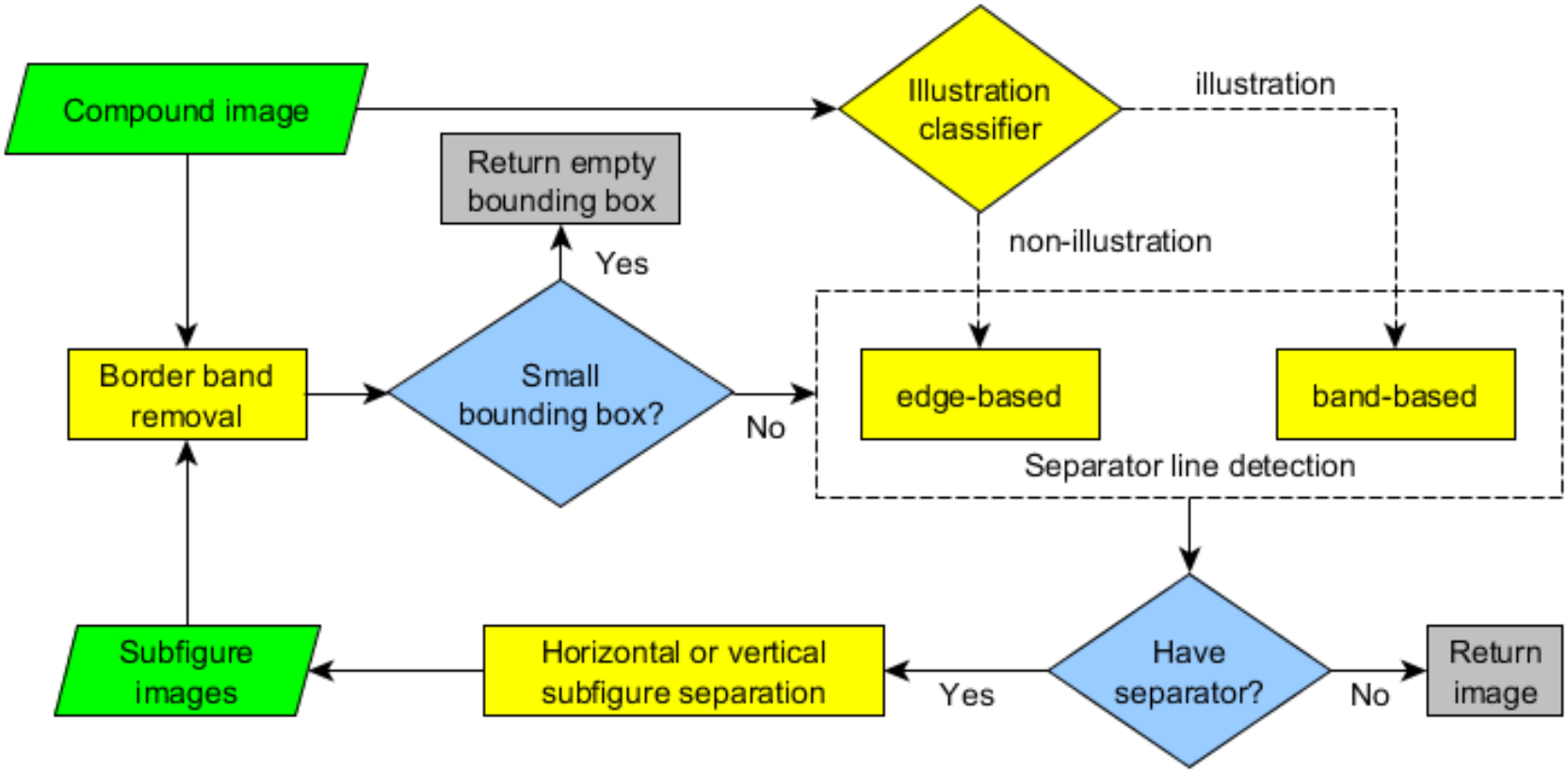
- Most approaches detect separator bands
- Very few detect separator edges
- No automatic combination yet
- We propose automatic selection of edge-based / band-based separation

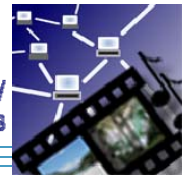




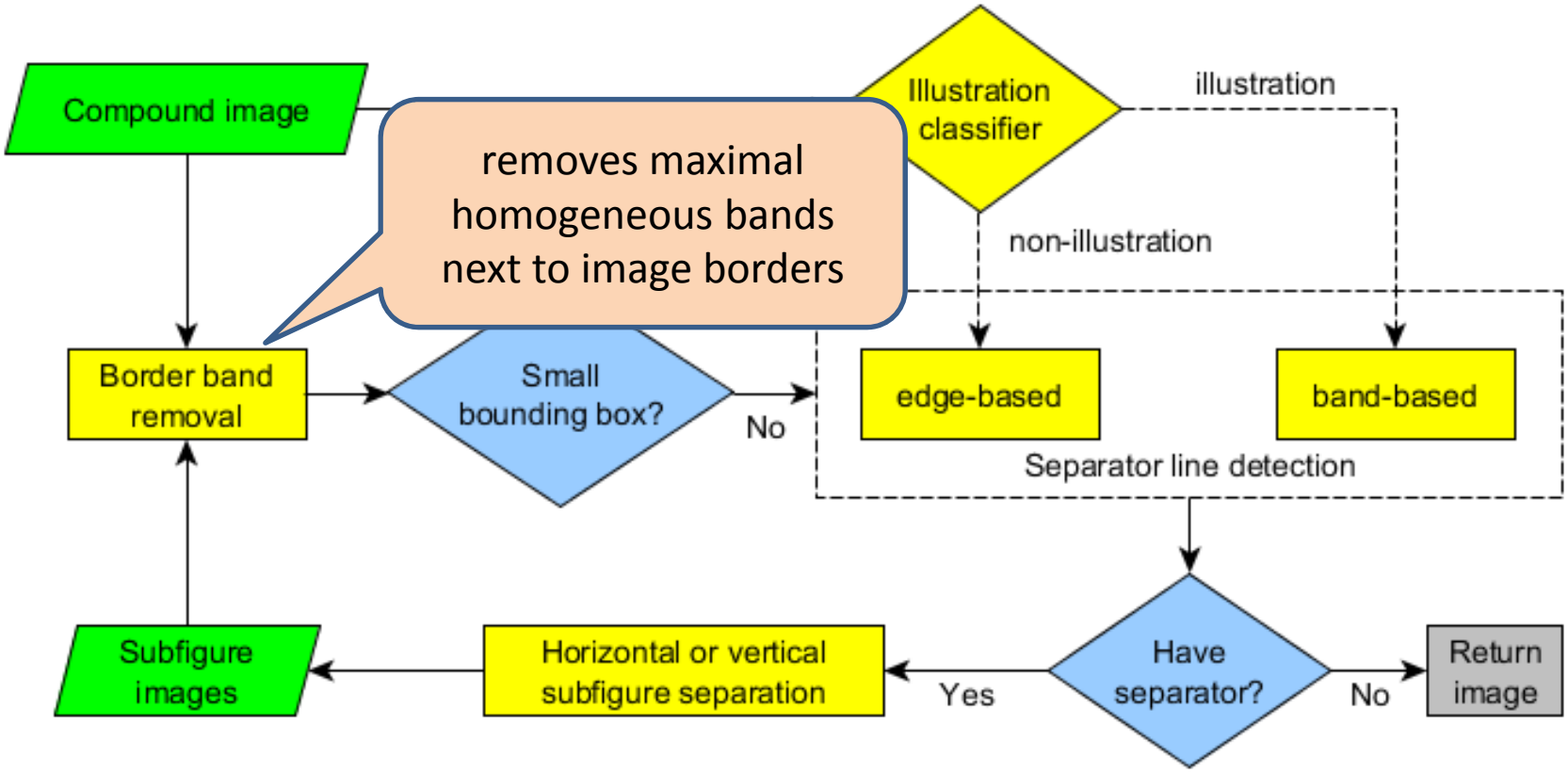


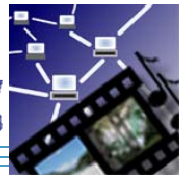
PROPOSED APPROACH



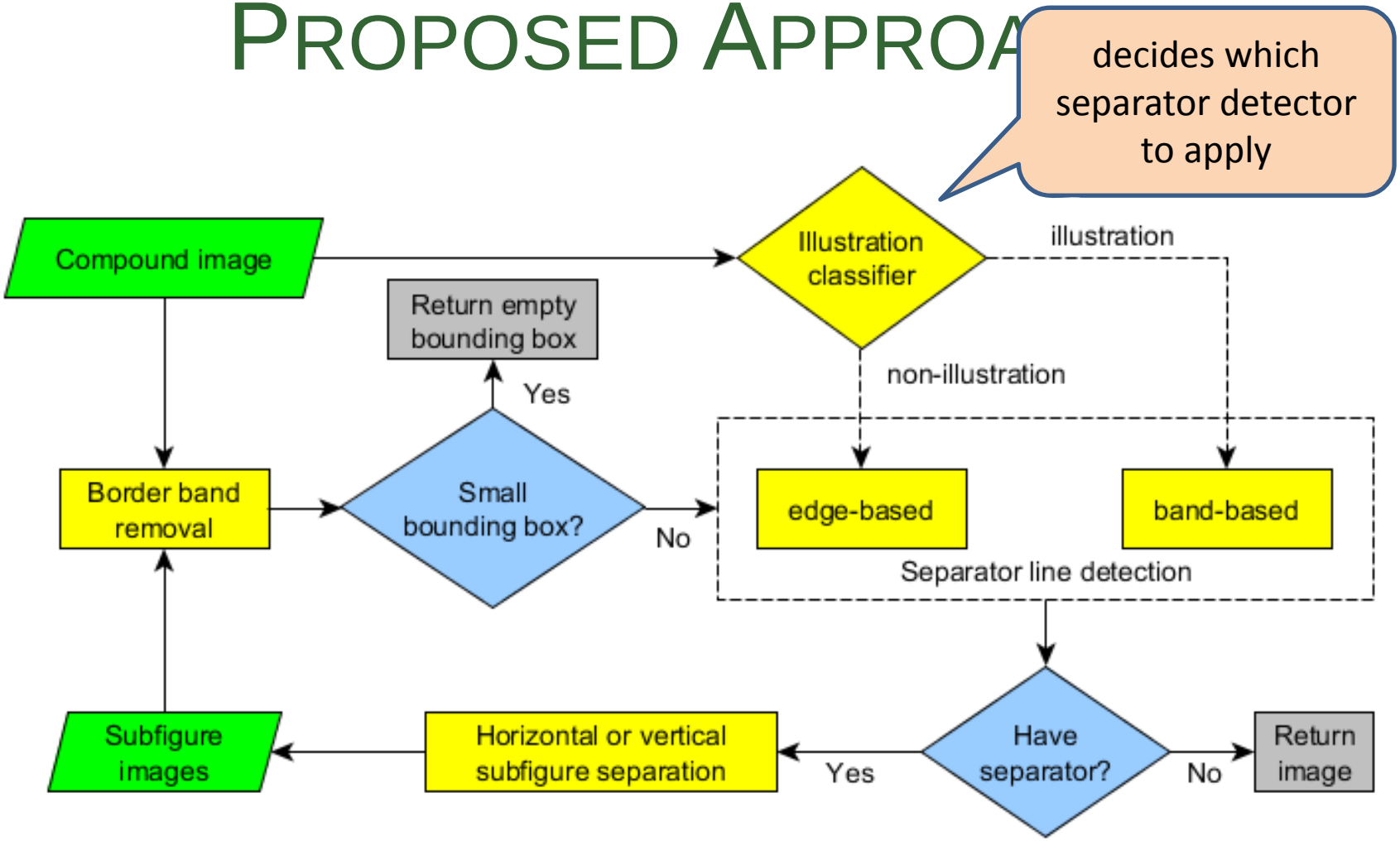


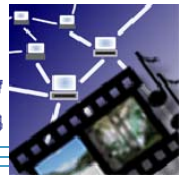
PROPOSED APPROACH



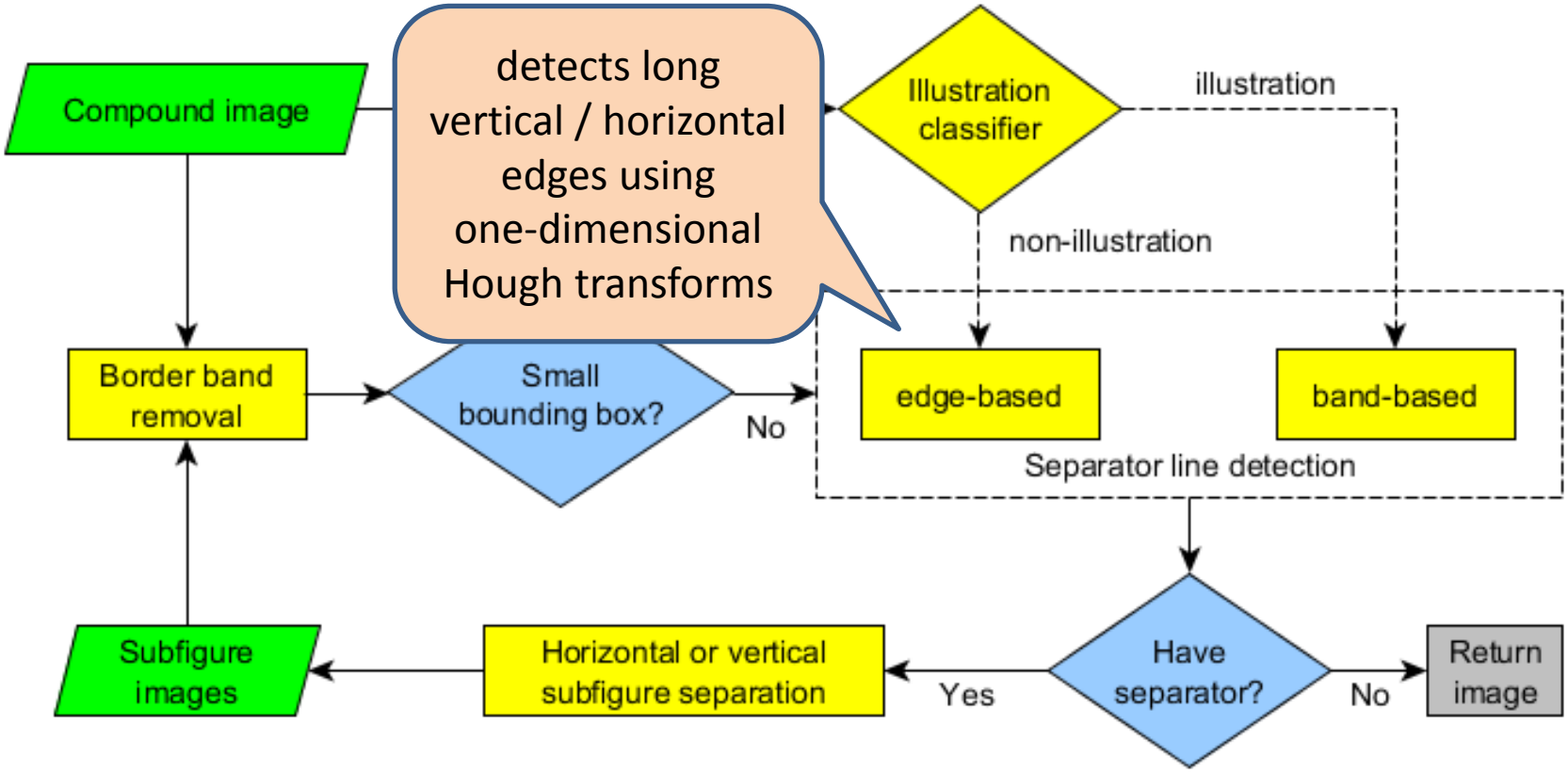


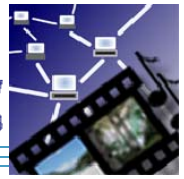
PROPOSED APPROA



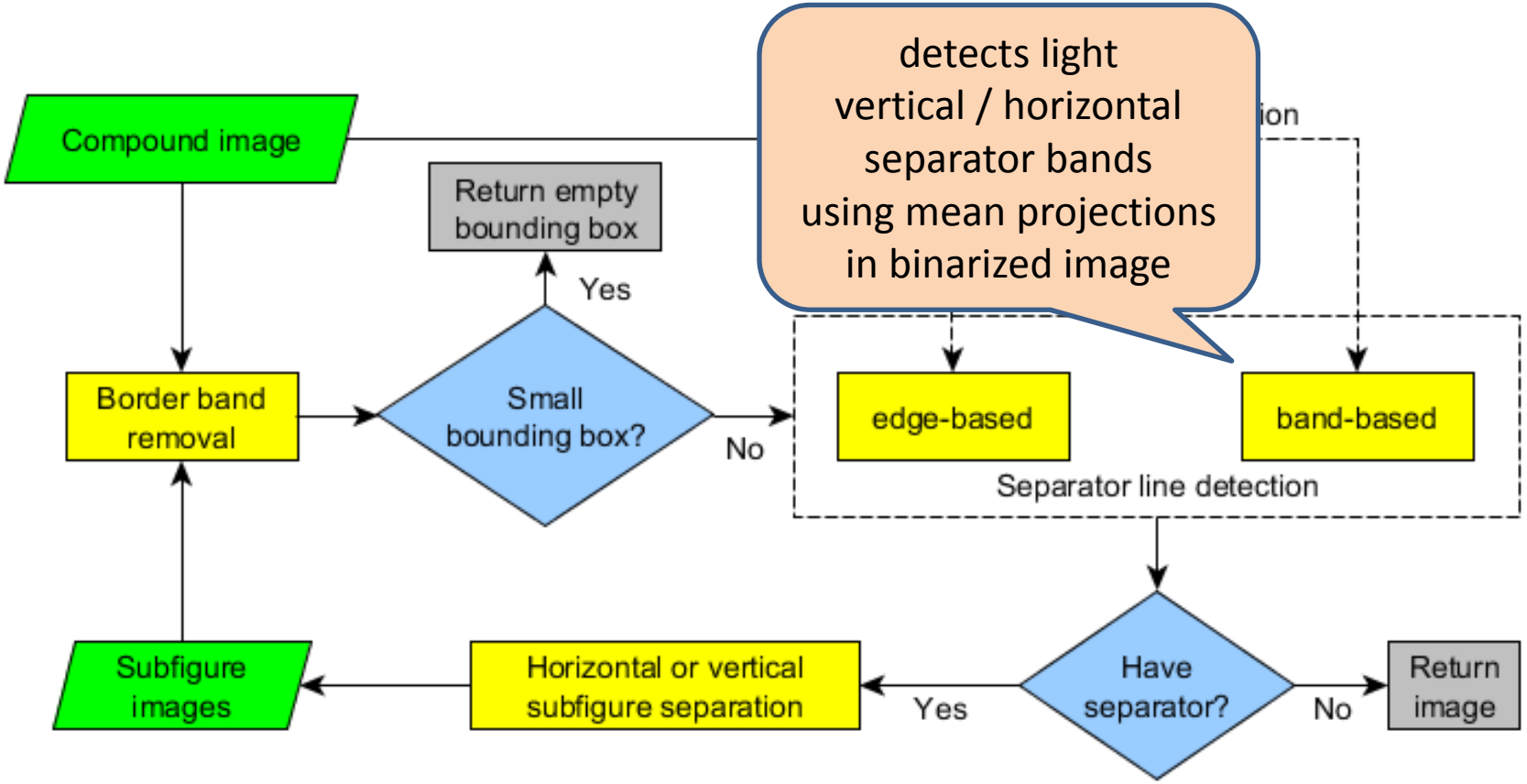


PROPOSED APPROACH

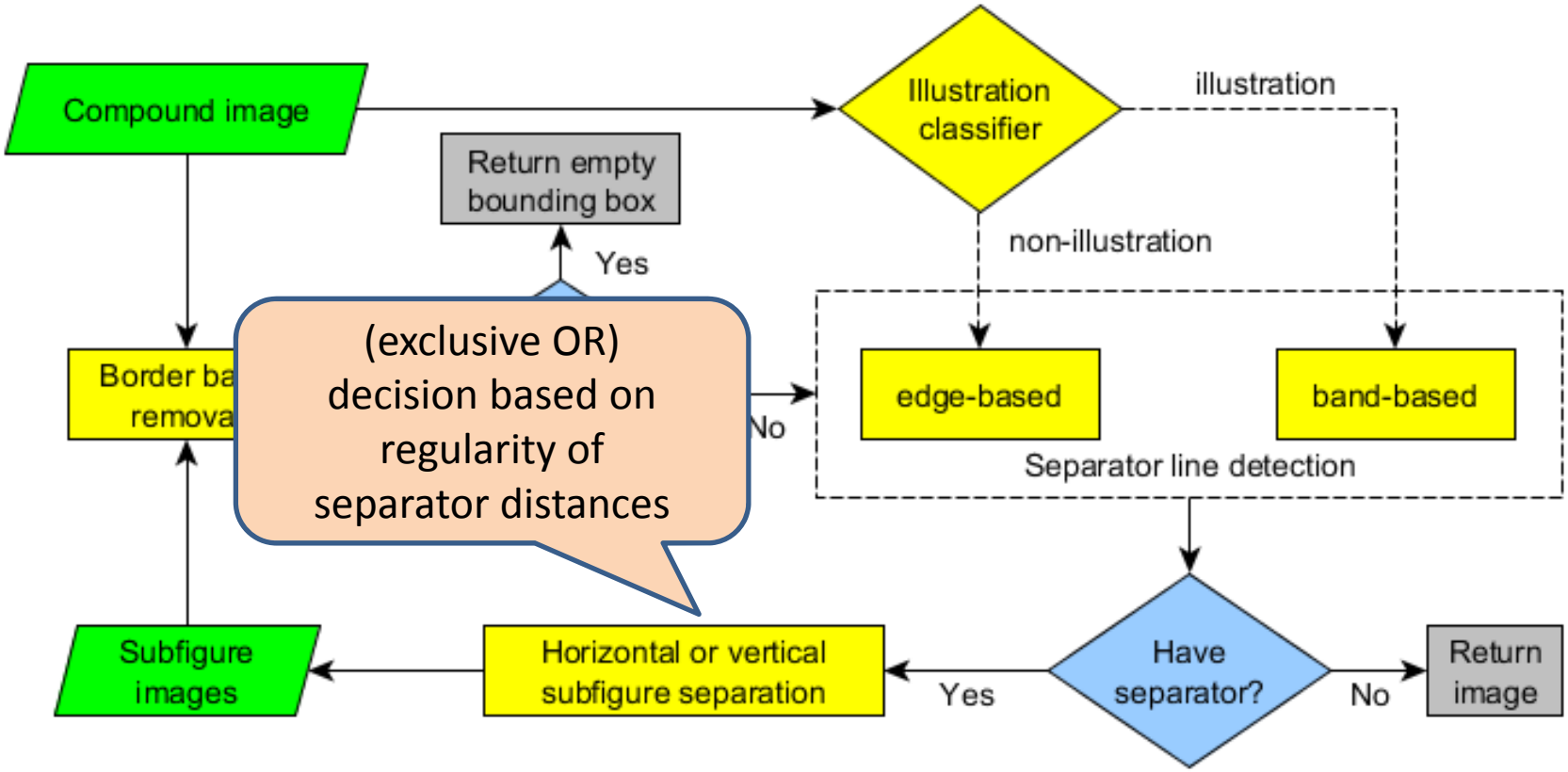




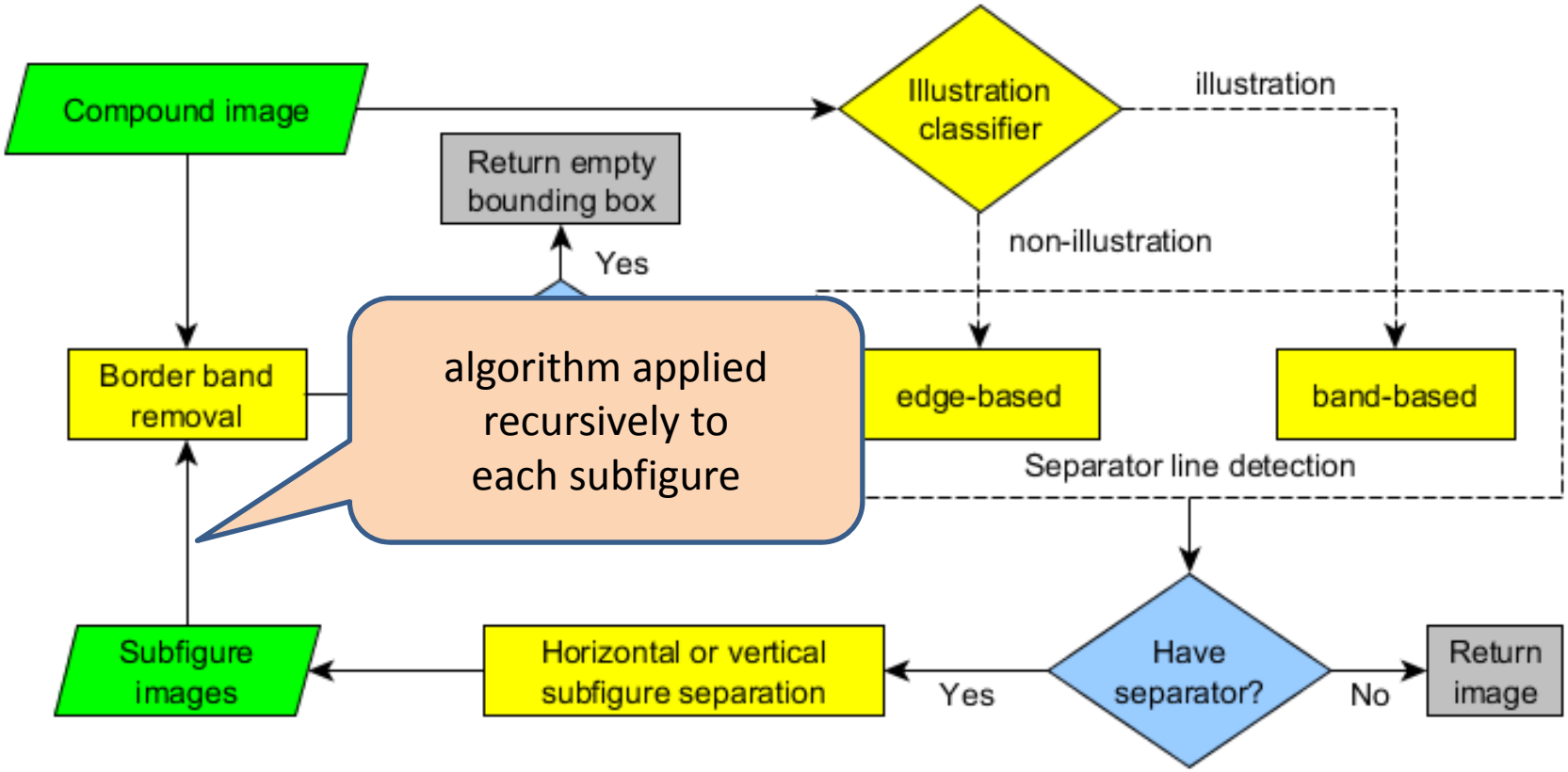
PROPOSED APPROACH



PROPOSED APPROACH



PROPOSED APPROACH

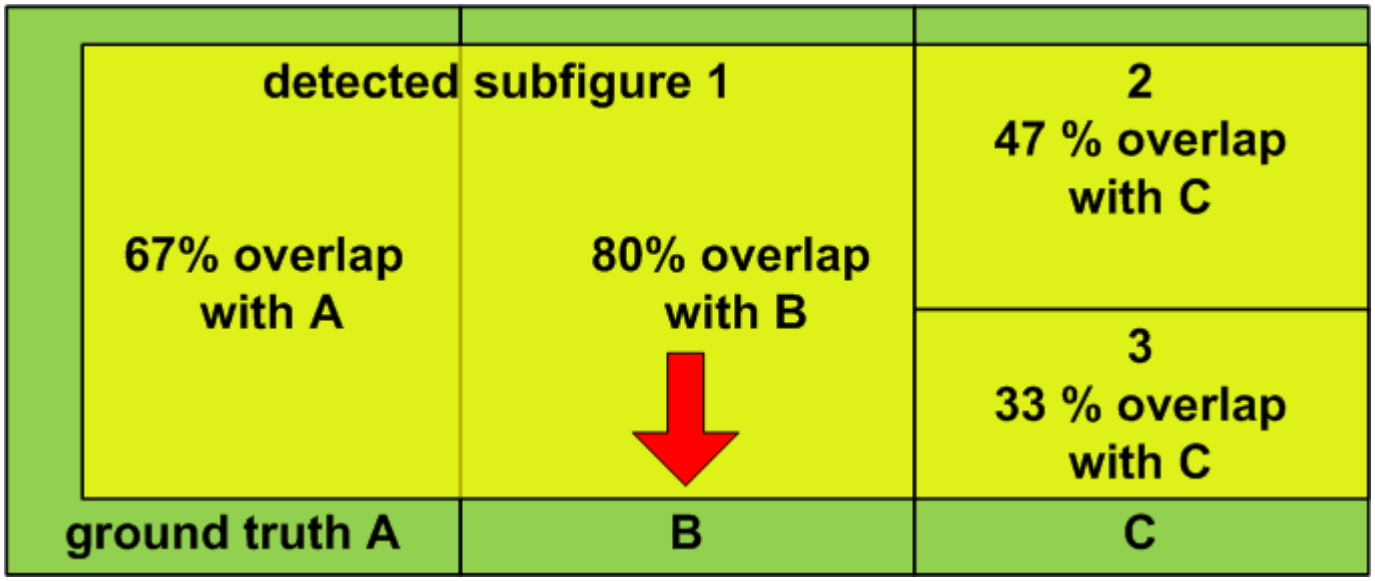


DATASET 1 FOR EVALUATION

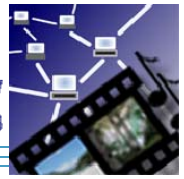
- ImageCLEF 2015 test dataset:
  - 3,381 compound figures from biomedical journals containing 12,789 ground-truth subfigures
- Accuracy is defined per compound figure:
  - N_D – number of detected subfigures
  - N_G – number of ground-truth subfigures
  - True positives TP: 1-to-1 mapping from detected to ground-truth subfigures (maximal overlap ≥ 66%)
  - Accuracy = TP / max(N_D, N_G)
- Report mean accuracy on test dataset

ACCURACY ON DATASET 1

- N_D – number of detected subfigures
- N_G – number of ground-truth subfigures
- True positives TP: 1-to-1 mapping from detected to ground-truth subfigures (maximal overlap ≥ 66%)
- Accuracy = TP / max(N_D, N_G)



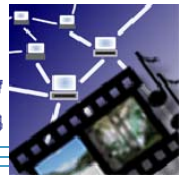
$N_D = N_G = 3$
TP = 1
Accuracy = 1/3



RESULTS ON DATASET 1

Method	Classifier / Features	Band-based %	Accuracy %
Proposed	None	0	58.0
Proposed	None	100	82.2
Proposed	SVM / simple11	60.3	83.5
Proposed	LogReg / simple11	74.1	**84.9**
NLM [7]	Manual	95.7	84.6

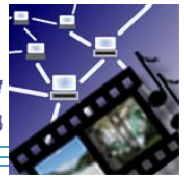
- LogReg: logistic regression, predicts class probability
  - decision threshold optimized on CFS training set
- simple11: 11-dimensional global image feature
  - entropy, mean intensity
  - 9 quantiles of intensity distribution



RESULTS ON DATASET 1

Method	Classifier / Features	Band-based %	Accuracy %
Proposed	None	0	58.0
Proposed	None	100	82.2
Proposed	SVM / simple11	60.3	83.5
Proposed	LogReg / simple11	74.1	**84.9**
NLM [7]	**Manual**	95.7	84.6

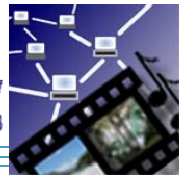
- Proposed approach outperforms
  - proposed variants without illustration classifier
  - semi-automatic approach of U.S. National Library of Medicine (NLM, best submission at ImageCLEF 2015)



RESULTS ON DATASET 1

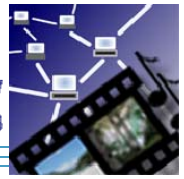
Method	Classifier / Features	Band-based %	Accuracy %
Proposed	None	0	58.0
Proposed	None	100	82.2
Proposed	SVM / simple11	60.3	83.5
Proposed	LogReg / simple11	74.1	**84.9**
NLM [7]	Manual	95.7	84.6

- Dataset is “biased” towards separator bands
  - NLM’s manual classification identified 96% band-separated compound figures
  - explains why our band-based-only variant achieves good performance



Dataset 2 for Evaluation

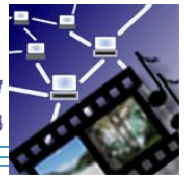
- NLM dataset [1]:
  - 389 compound figures from biomedical domain containing 1,754 ground-truth subfigures
- Stronger criterion for true positive subfigures:
  - ≥ 75% overlap with a single ground-truth subfigure
  - < 5% overlap with all other ground-truth subfigures
- Precision, recall and F_1 measure
  - calculated from total numbers of **detected**, **true positive** and **ground-truth** subfigures on entire test dataset



RESULTS ON DATASET 2

Method	detected	TP	Precision %	Recall %	F1 %
Proposed, SVM/simple11	1681	1392	82.8	79.4	81.1
Proposed, LogReg/simple11	1646	1407	85.5	**80.2**	**82.8**
NLM [1]	1482	1276	**86.1**	72.3	78.6

- Indicate generalization capability:
  - used same parameter settings as with dataset 1
  - relative performance consistent with previous results
- Band-based separator selection rate: 33%
  - substantial difference to dataset 1 (74%)



CONCLUSION AND FURTHER WORK

- Proposed compound figure separation approach:
  - uses a **supervised classifier** to select separator line detection method (band-based or edge-based)
    - classifier accuracy is not critical for CFS performance
    - optimizing classifier's decision on CFS training set helps
    - future work may include finding more discriminative features / better training sets
  - consistently **better than previously published results** on 2 datasets, using same parameter settings
  - may be **extended by other known useful techniques** (image markup removal, subfigure label recognition)