Fachhochschule
Bonn-Rhein-Sieg

Human Detection and Action Recognition in Video Sequences

Human Character Recognition in TV-Style Movies

Alexander Kläser

Cordelia Schmid – LEAR, INRIA Rhônes Alpes
Rainer Herpers – Fachhochschule Bonn-Rhein-Sieg

6. December 2006

Overview

- Introduction: goal, sample images, overview approach
- Approach: human detection, character identification
- Results: images, videos
- Closure: conclusion, possible extensions, acknowledgements

Overview

Goal

- Detecting humans in video sequences and determine their identity
- Type of data: TV-/cinema-style videos, they provide an uncontrolled, realistic working environment
- Motivation
  - Understand image content
  - Applications such as image/video retrieval, automated video surveillance

Example Images




Approach Overview

(1) Segment the video into shot sequences (data structuring)

(2) Human detection (supervised learning)

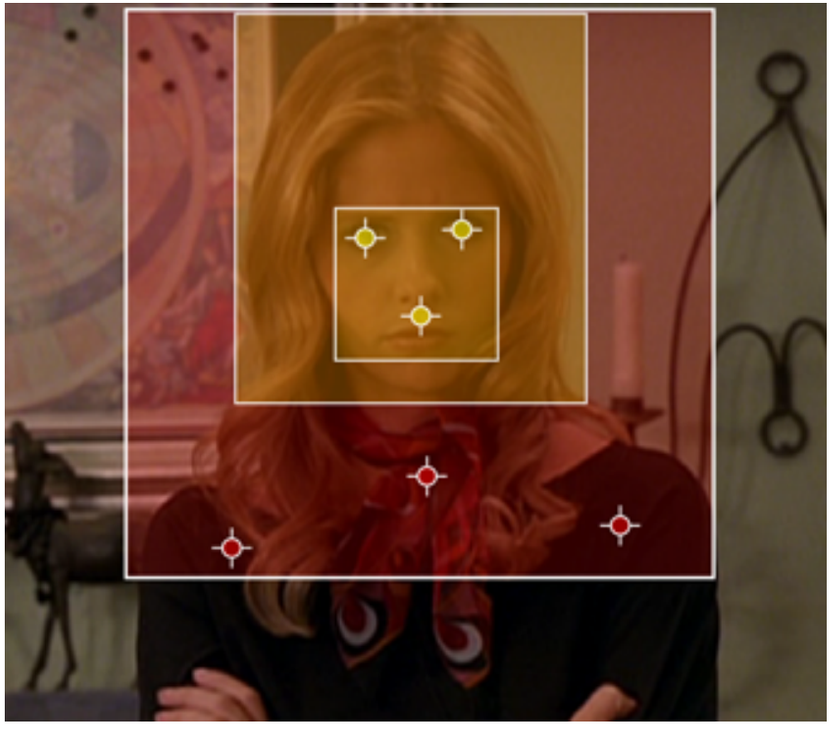
(3) Character identification (supervised learning)

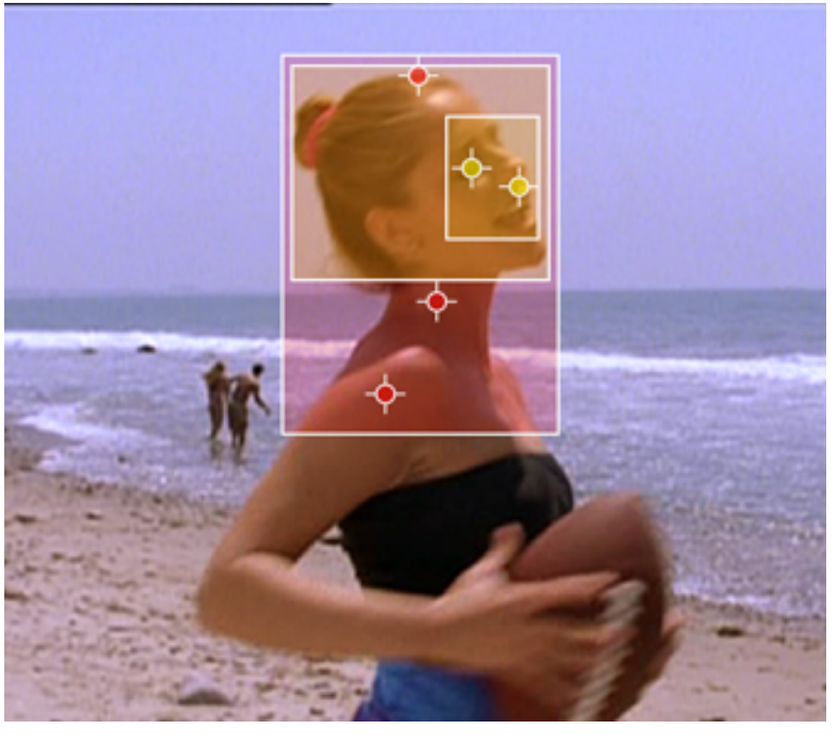
Human Detection

Goal

- Human detection in images under the following constraints:
  - Lateral/frontal
  - From closeup to distant view
  - Partial occlusion
- Approach:
  - Detection of body parts (face, head, head+shoulders)
  - Flexible assembly of detected body parts
  - Gaussian model for geometric relation between body parts

Annotation Examples





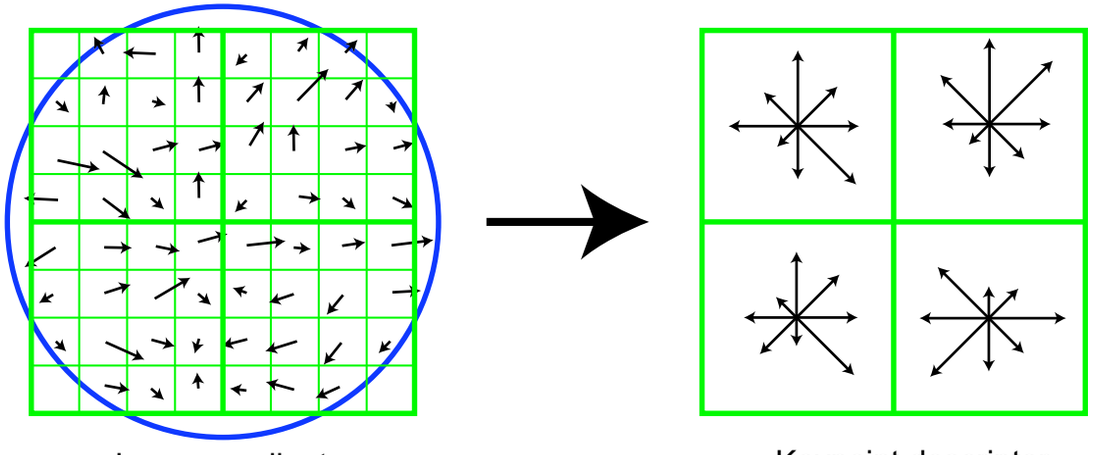
Body Part Detection

- Detector introduced by Dalal and Triggs [1] is trained on body parts
  - Detector computes SIFT-like descriptors on a fixed grid and uses SVMs for learning

  $\Rightarrow$ Characteristic structures (esp. edges in images) are abstracted and learned

- Altogether 6 different detectors: face, head, head+shoulders, each frontal+lateral

SIFT-like? – The SIFT-Descriptor [3]



(1) An image region is divided into cells (here 2×2)

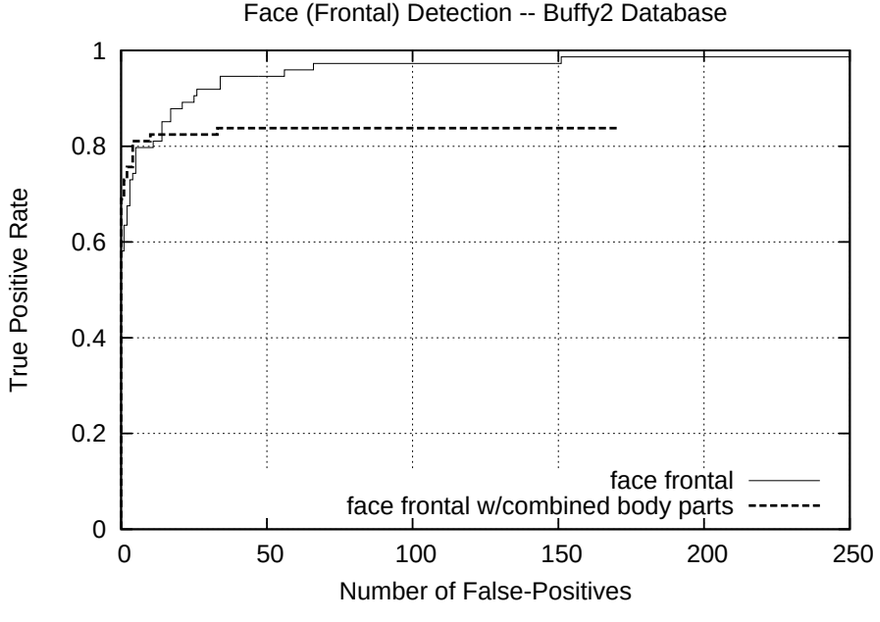
(2) Gradient orientation and magnitude are computed for each pixel in a cell (here 4×4 pixel)

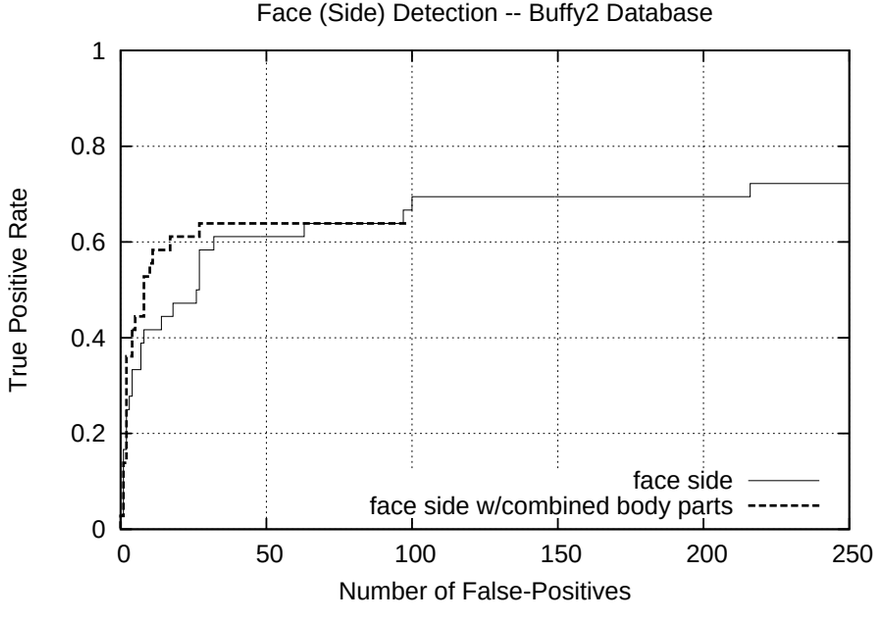
(3) All pixels (weighted with a Gaussian) vote into the Histogram of Oriented Gradients (HoG) of their cell

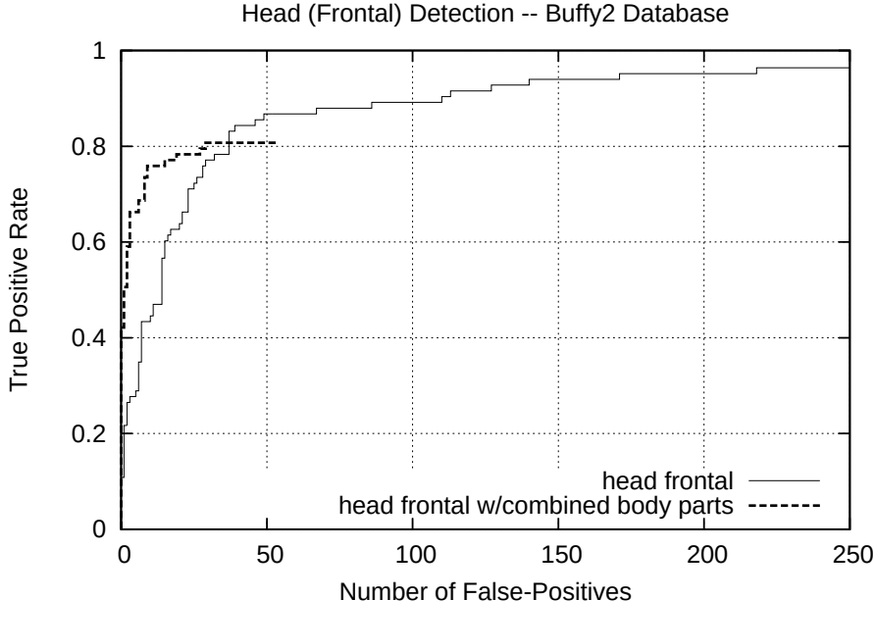
Geometric Relation between Body Parts

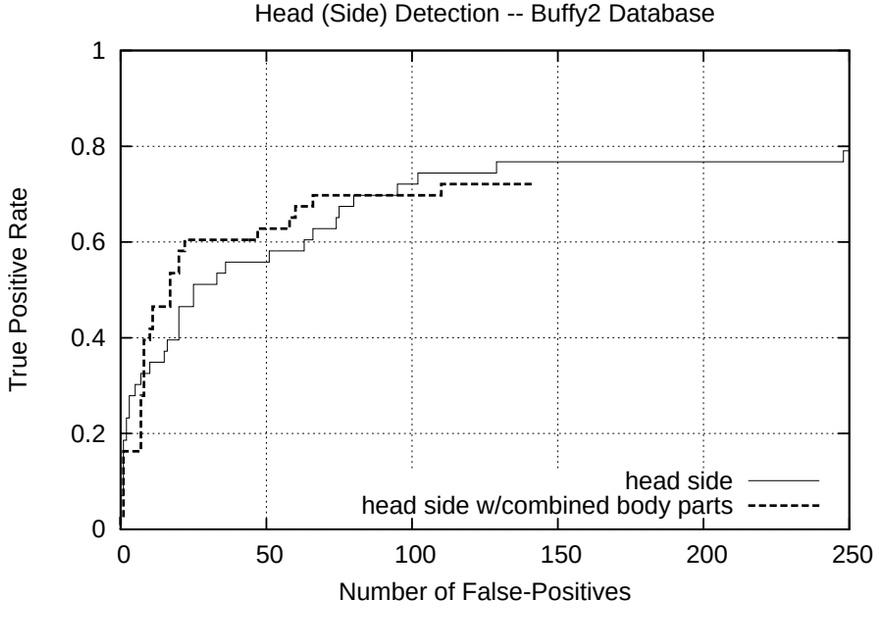
- Model similar to work of Mikolajczyk und Schmid [2]
- On training data and for each body part combination, Gauss functions are learned for: relative x-/y-position, relative scale
- Detected body parts are assembled based on this model
- Yields a more robust detection

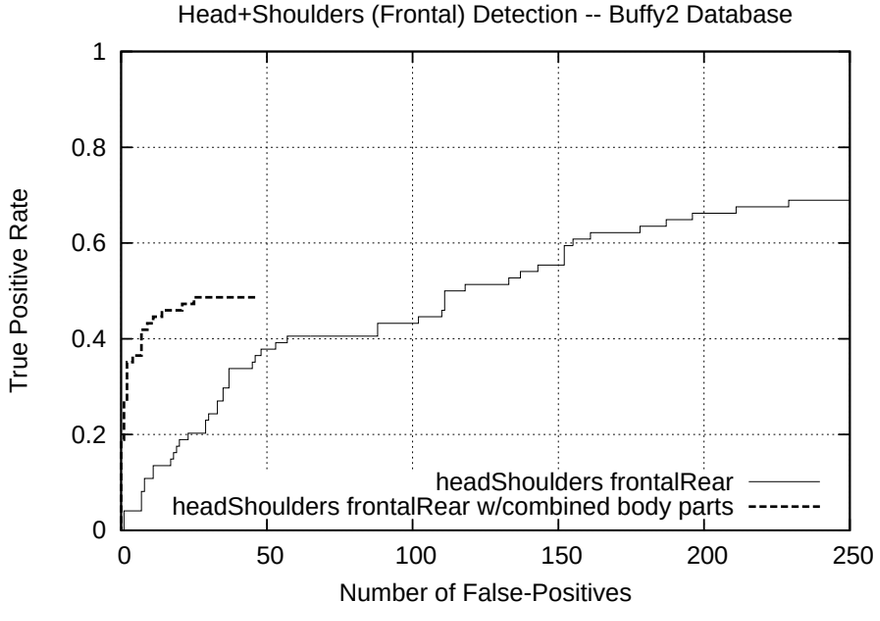
Statistics

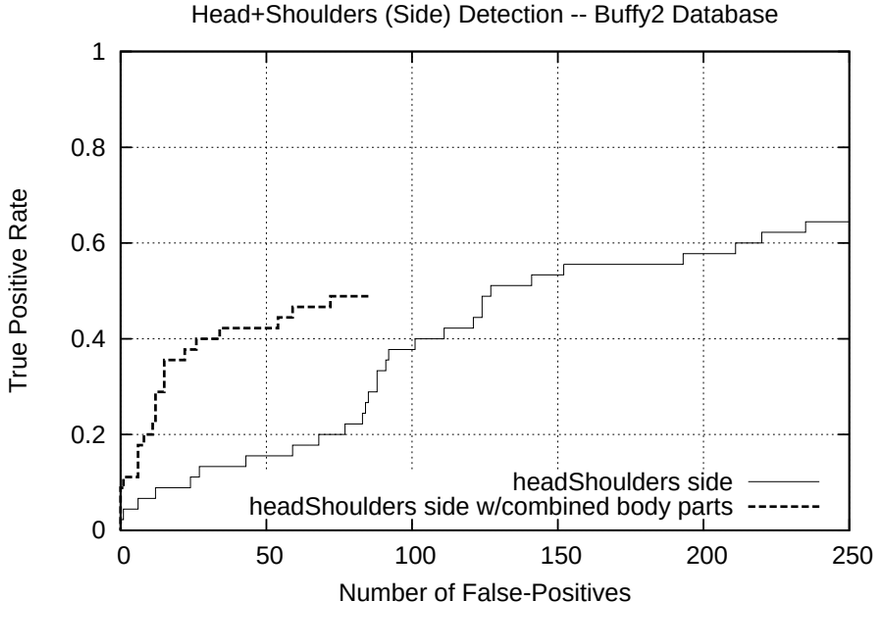

Face (Frontal) Detection -- Buffy2 Database
True Positive Rate
Number of False-Positives
face frontal
face frontal w/combined body parts
0
0.2
0.4
0.6
0.8
1
0
50
100
150
200
250


Face (Side) Detection -- Buffy2 Database
True Positive Rate
Number of False-Positives
face side
face side w/combined body parts
0
0.2
0.4
0.6
0.8
1
0
50
100
150
200
250


Head (Frontal) Detection -- Buffy2 Database
True Positive Rate
Number of False-Positives
head frontal
head frontal w/combined body parts
0
0.2
0.4
0.6
0.8
1
0
50
100
150
200
250


Head (Side) Detection -- Buffy2 Database
True Positive Rate
Number of False-Positives
head side
head side w/combined body parts
0
0.2
0.4
0.6
0.8
1
0
50
100
150
200
250


Head+Shoulders (Frontal) Detection -- Buffy2 Database
True Positive Rate
Number of False-Positives
0
0.2
0.4
0.6
0.8
1
0
50
100
150
200
250
headShoulders frontalRear
headShoulders frontalRear w/combined body parts


Head+Shoulders (Side) Detection -- Buffy2 Database
True Positive Rate
Number of False-Positives
0
0.2
0.4
0.6
0.8
1
0
50
100
150
200
250
headShoulders side
headShoulders side w/combined body parts

Character Identification

Goal

- Determine the identity of a person under the following constraints:
  - Varying poses, varying perspectives
  - A person can change clothes in a video sequence
  - Varying environments, varying illumination conditions
- Approach:
  - Bag-of-Features (BoF) separately for each detected body part
  - Classes for main characters + all others
  - Combined codebooks: SIFT and color
  - More robust identification by combining probabilistic votes of connected body parts

Example Training Images (from Buffy)


face
head
head+shoulders
frontal
side
frontal
side
frontal
side
Buffy
Giles
Riley

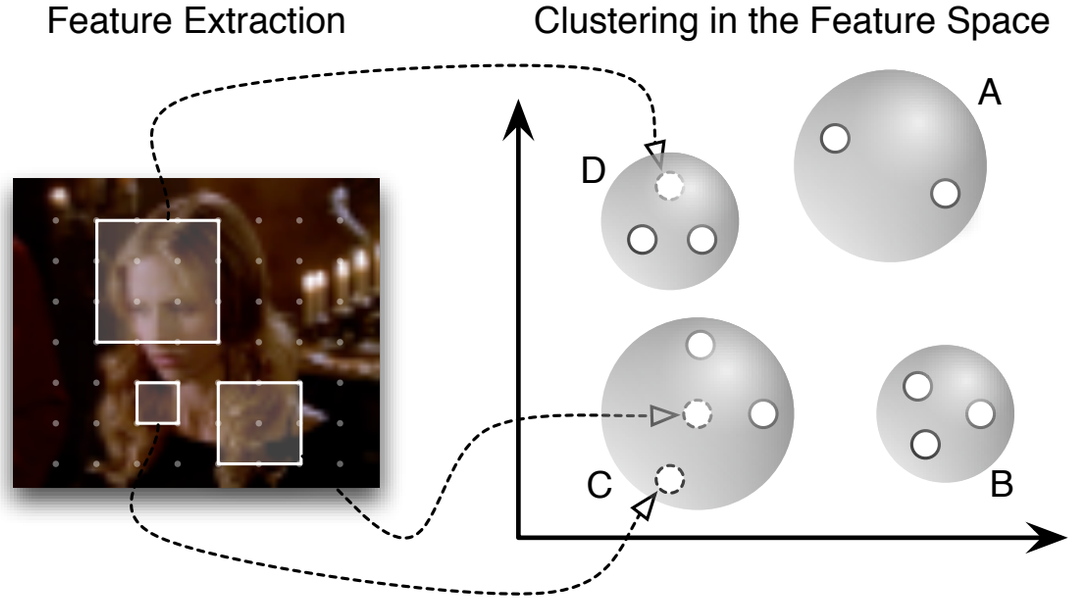
Bag of Features [4, 5]

(1) Feature extraction

- Feature: abstracts the local neighborhood of a point in an image (we use SIFT, and mean CIEL*U*V* color)
- Given: training images for different body parts (face, head, head+shoulders) and classes (main characters + others)
- Features are extracted from *all* images (of certain body part type) using dense sampling

(2) Codebook generation

- All features can be seen as a distribution in a high-dimensional feature space
- A cluster algorithm (we use k-means) groups similar features together into *clusters*
- Each cluster corresponds to a *visual word*, alle words together represent the *codebook*

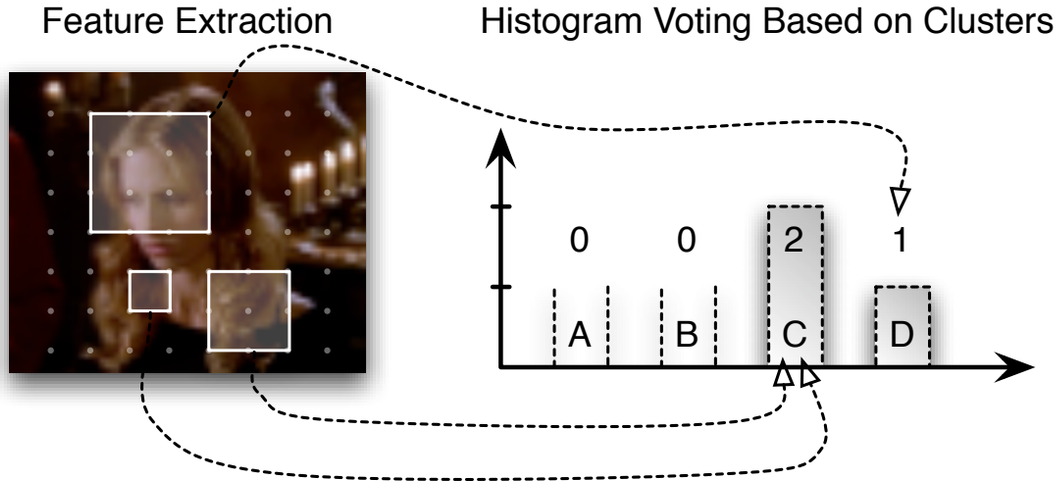

Feature Extraction
Clustering in the Feature Space
A
D
C
B

(3) Classification

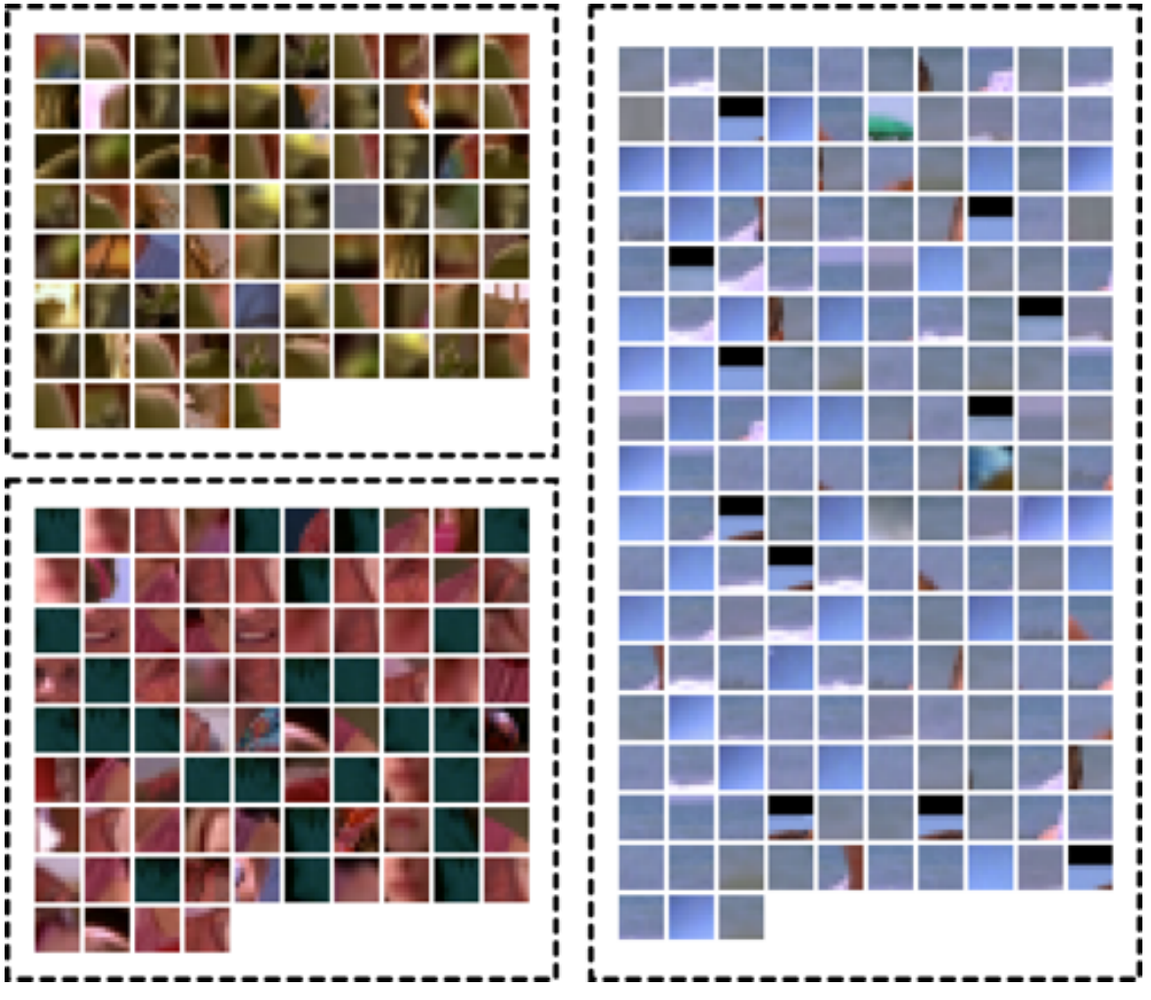
- Features are extracted (in the same manner as before) from *one specific* training image
- Each feature is assigned to its closest visual word

$\Rightarrow$ All features of an image yield a word histogram (we binarize it)

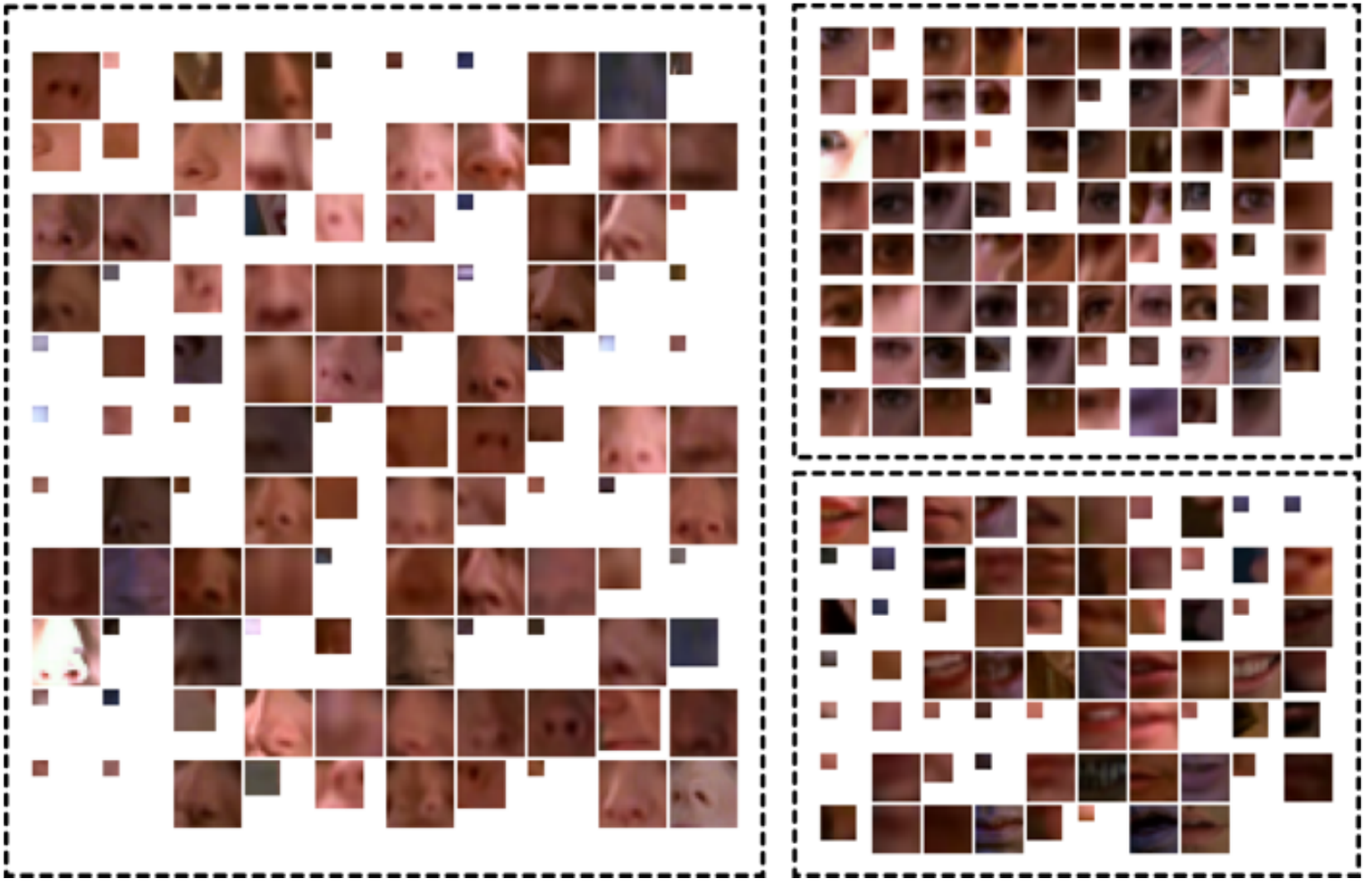
- Based on the histograms, an SVM learns (and predicts) the class membership = identity of a person (one-against-one multi-class SVM, non-linear with RBF kernel)


Feature Extraction
Histogram Voting Based on Clusters
0
0
2
1
A
B
C
D

Cluster Examples (Color)



Cluster Examples (SIFT)



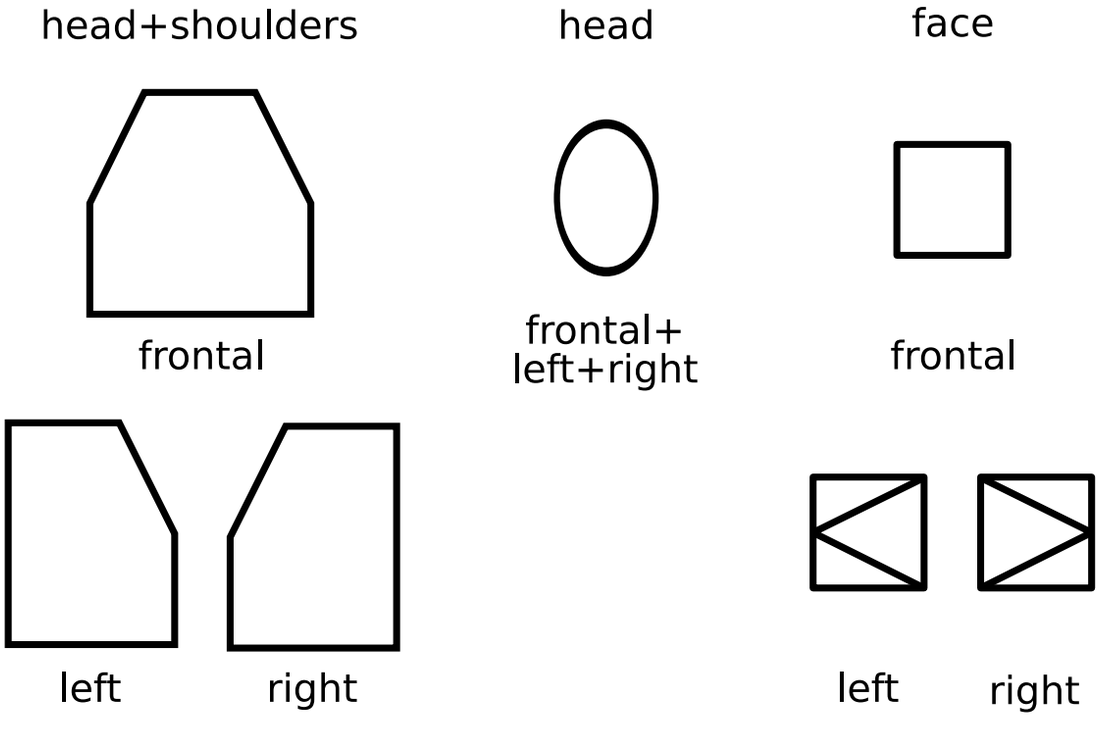
Combination of Body Parts

- The character identification is more robust if results of connected body parts are combined
- For each detected body part, a probabilistic vote (containing the membership probability for each class) is computed
- Combination: mean from all votes, the class with the highest probability wins

$\Rightarrow$ Accuracy over all classes and all body parts 81.3% w/o combination 74.4%) (on Buffy data set, leave-one-out cross validation)

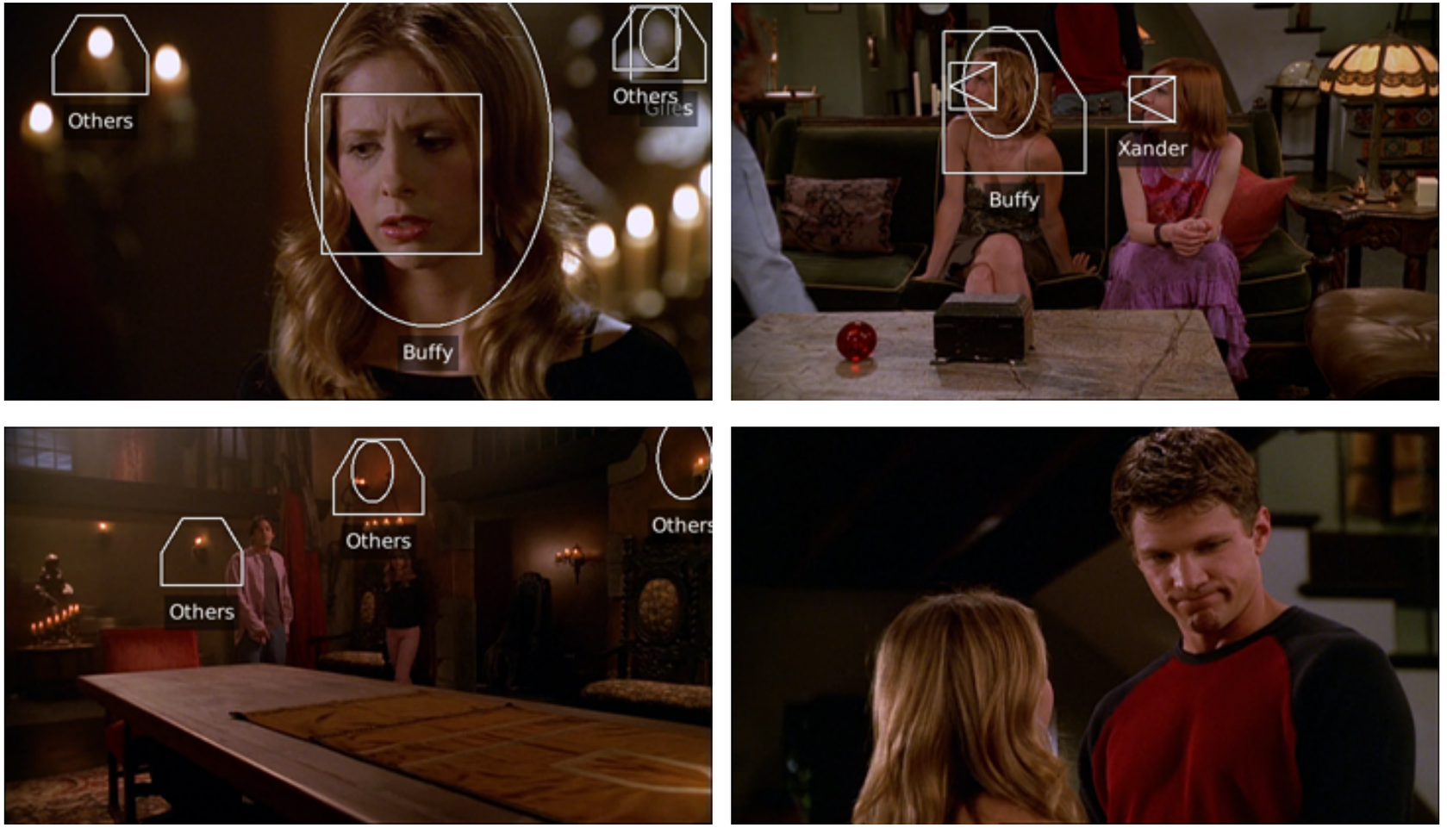
Results

Result Visualization



Results (correct)

Results (incorrect)



Closure

Conclusion

- A system for human detection and character identification in TV-/cinema-style videos has been successfully developed, implemented, and tested
- Approach based on supervised learning methods
- Two main techniques:
  - Human detection combines detection of body parts
  - Character identification uses a bag-of-features approach
- Very promising results
- Combination of body parts improves results significantly

Future Work

- Human Detection
  - Different detection systems for body parts (esp. for body parts that are less rigid), e.g. based on contours
  - Use temporal information in the video sequence
  - Additional tracking of detected body parts
- Character Identification
  - Different, more versatile clustering methods
  - No binary word histograms
  - Feature extraction using e.g. interest point detectors

  - Different feature descriptors esp. for color (e.g. Color-SIFT), but also in general (e.g. SURF)
  - Extension for unsupervised learning

- Towards image understanding
  - Scenes and environments could be detected with a BoF-like approach
  - Recognized scenes can help to reason on a more semantic level (e.g. about the interaction between certain characters)

References

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886-893, June 2005.

[2] Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*, volume I, pages 69-81, 2004.

[3] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91-110, 2004.

[4] Shivani Agarwal and Dan Roth. Learning a sparse representation for object detection. In *Proceedings of the 7th European Conference on Computer*

Vision, volume 4, pages 113-130, 2002.

[5] Chris Dance, Jutta Willamowski, Lixin Fan, Cedric Bray, and Gabriela Csurka. Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

So far about my master thesis. . .

. . . a big THANKS to the group :) . . .

. . . do you have questions?