Human Detection and Character Recognition in TV-Style Movies

Alexander Kläser

Human Detection and Character Recognition in TV-Style Movies

Alexander Kläser
Fachhochschule Bonn Rhein Sieg
alex.klaeser@gmx.de

1 Objective

The objective of this master thesis is the recognition of human characters in TV-style video sequences. In order to recognize the same person at different time instances in a video sequence, the outward appearance of the person has to be described and learned with an appropriate model. The diversity in which humans can appear makes the task of human detection and character recognition to a particularly challenging problem. TV-style movies provide an uncontrolled, realistic working environment for human detection and character recognition. Possible applications range from surveillance (e.g., intrusion detection) and security applications (e.g., person identification) to image retrieval or semi-automatic image annotation (e.g., automatic labeling of faces in personal photo albums).

2 Approach

A supervised approach is proposed that can be split into three stages. After a pre-processing stage, humans are detected at different time instances in a given movie. Then, the identity of each detected person is computed based on learned models. A more detailed overview on the different stages is given in the following.

Pre-processing. Extracted full-frames of video sequences are considered since no temporal information is included. In order to structure the large amount of data, the extracted frames are grouped into shots. The shot detection is based on the comparison of color histograms of successive frames.

Human Detection. The human detection combines detectors of upper body parts which are *face*, *head*, and *head+shoulders*. They are divided into *frontal* and *side* view (i.e., six detectors in total). The body part (view) detector in this work is based on the approach introduced by Dalal and Triggs [DT05]. Their system samples densely overlapping SIFT-like [Low04] feature points which are trained with a linear Support Vector Machine.

In order to obtain more robust results, single body part detections are combined applying knowledge about geometric relations between the body parts. This is done by learning a Gaussian model similar to the work of Mikolajczyk et al. [MSZ04].

In this way, humans are detected in varying views (frontal and side) and in resolutions varying from distant up to close-up views. Partly occluded people are detected as well. For the training of the body part detectors, an annotation data set is created that consists of roughly 330 training annotations for each body part (of which about $\frac{1}{3}$ are for the side and $\frac{2}{3}$ for the frontal view).

Human Character Recognition. Our approach for the character recognition was motivated by the *Bag-of-Features* method [DWF$^+$04, AR02] that has been used in previous work for object classification tasks. The Bag-of-Features method extracts feature points (i.e., image points that are described not necessarily by their color/intensity values, but by their local neighborhood based on, e.g., gradient information) from a set of training images. In their feature space, the feature points are grouped by a clustering algorithm. Based on the resulting clusters (all clusters together are referred to as *code book* and one cluster is referred to as *visual word*), occurrence histograms are generated for each body part image. A classifier is then trained on these histograms. Occurrence histograms reflect how many feature points are assigned to each visual word.

Our approach is build on SIFT- [Low04] and CIE L*u*v* color-based code books that are obtained by clustering with k-means. A non-linear multi-class Support Vector Machine (SVM) is learned on occurrence histograms for five main characters and on one category "Others" that consists of all other characters. The trained Support Vector Machines (or SVM models) are then used to predict the identity of a detected person. Probabilistic votes of connected body parts (i.e., body parts that belong to one and the same person) are combined for a more stable prediction. The training data is generated from an annotation data set in which the name of the corresponding character is noted for each body part. Since a supervised approach is applied and since the focus lies on learning character identities for a particular video (e.g., a complete TV/cinema movie), example annotations for this particular movie are created. Based on these annotation data, codebooks are generated and SVM models are learned. The codebooks and the SVM models are then applied subsequently on the entire video sequence.

In this way, particular (human) characters are recognized at different points in time in a given video sequence. Our method takes into account that humans change in pose, perspective, and distance to the camera, that illumination conditions change, and the same character can wear different clothes.

3 Results

Figure 1 shows performance plots for the different body part detectors. The plots show the ROC curve for each single detector as well as for each detector in combination with the other detectors, i.e., employing the model for the geometric relations of the body parts.

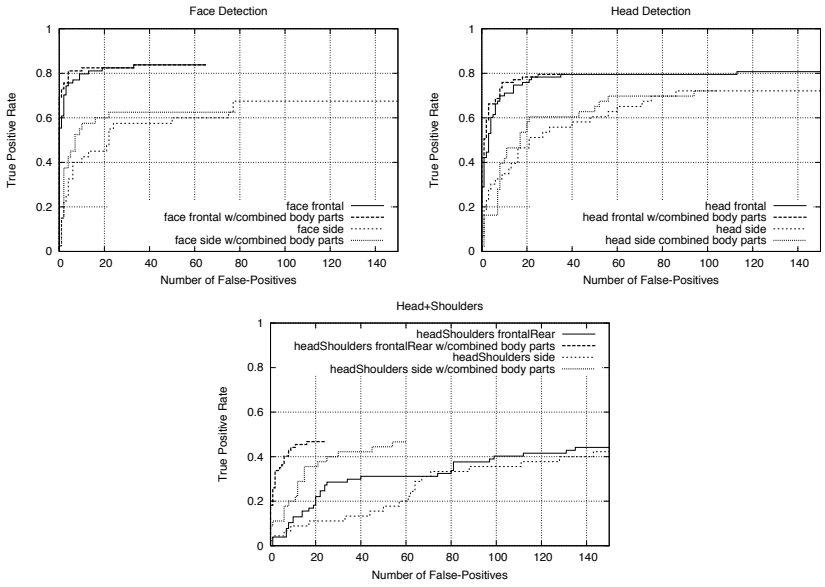

Figure 1: The performance of the (single as well as combined) body part detectors plotted as ROC curve with the number of false positives.

The frontal face detector performs best. It can be seen that the combination of the different detectors can always improve the performance, especially for detectors with an initial low performance.

Figure 2 shows some examples of typical correct detections and character labelings. As example video sequence, the first episode of *Buffy the Vampire Slayer Season 5* has been used. Different orientations and combinations of the body parts are detected successfully (see Figure 2 bottom right, direction is indicated through arrows in the face detections). If the size of a person in an image is too small, his faces cannot be detected. However, in order to detect the person, the detection of head and head+shoulders suffices (see Figure 2 bottom left). Partly occluded and overlapping people are detected correctly as well (see Figure 2 top row).

In the experiments carried out on our annotation set, the character recognition accuracy achieved 81.3% (over all character classes and all body parts using leave-one-out cross validation). For the training of the BoF-SVMs ca. 90 example annotations for each body part have been used for the class “Others” and for the main character Buffy, and ca. 40 examples have been used for the remaining four main characters. Among the cases where the character identification failed, mislabeling of other characters as belonging to one of the main character classes were more common than vice versa. Characters with for example blond hairs were several times classified as Buffy, whereas Buffy herself would be identified correctly in most cases. This is certainly due to the general nature of the class “Others”.

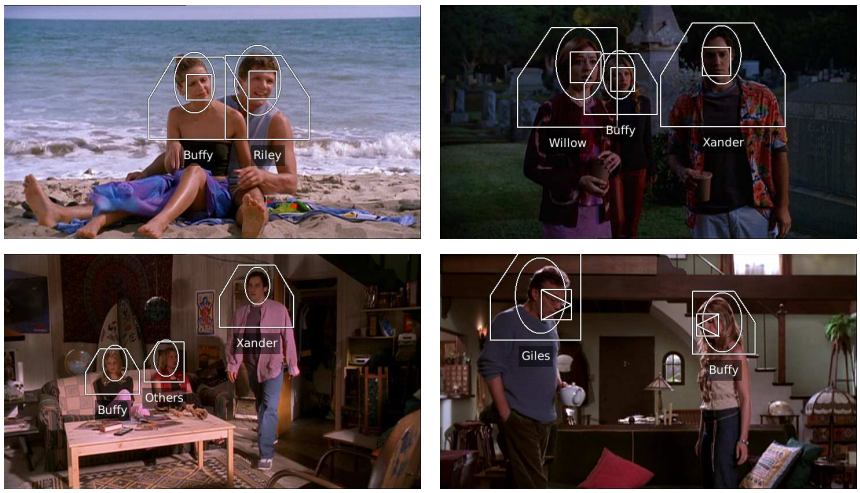

Figure 2: Typical examples for correct final detections and character identifications.

Conclusion. Good detection and recognition results were obtained with a system that combines different body part detectors and a Bag-of-Features based character recognition. Especially the combination of different body parts helps to improve the overall human detection performance. The character identification module worked well on the test sequence, although less sophisticated methods for the clustering (k-means) and the feature extraction (dense sampling) has been employed. Again, the combination of the votes from the different body parts helps to improve the final results.

References

[AR02] 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.

[DT05] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, June 2005.

[DWF+04] 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.

[Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[MSZ04] 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.