Integration of Structural and Color Cues for Robust Hand Detection in Video Images.

— Extended Abstract —

Jan Richarz, Thomas Plötz and Gernot A. Fink
Dortmund University, Robotics Research Institute, Intelligent Systems Group
Otto-Hahn-Str. 8, 44227 Dortmund, Germany
{Jan.Richarz,Thomas.Ploetz,Gernot.Fink}@udo.edu

Abstract

Image structure and color are those two cues that are most frequently used in image interpretation and computer vision tasks. However, both have weaknesses if applied separately. Fusing the information of different cues allows for compensating the deficits of one of them by exploiting the advantages and attributes of others, leading to more robust results. Here we propose an approach to hand detection in video images of realistic scenes combining structural and color features. To describe image structure, we focus on the gradient orientation histograms extracted by the Scale Invariant Feature Transform (SIFT) as one variant of scale- and rotation-invariant salient region descriptors. Color information is evaluated by a skin color detector based on Gaussian Mixture Models (GMMs). We investigate different schemes for fusing these two different types of information, and present evaluation results for each of them using a large database of images recorded inside a smart house.

1 Introduction

Our ongoing research incorporates the development of an "Intelligent House" as an integration scenario for different applications in the fields of multi-modal Human-Machine-Interaction (HMI) and pattern recognition. But what makes us perceive a house – or any system – as "intelligent"? We consider an interaction partner as smart if we may interact with him in the way we normally would with other humans, and he shows reasonable reactions to our actions. So, an "intelligent" system is not only defined by the services it offers (however useful they may be), but also – and more importantly – by the naturalness of interfaces it offers to access these services. Consequently, we seek to design human-machine interfaces that resemble natural means of communication.

In our scenario, the user should be able to control different features of the house (like lighting and sun-blinds) using gestures and speech. For this purpose, the house is – among others – equipped with active cameras and microphones. Interaction should take place as natural as possible, which means we do not want to constrain the way gestures are performed or commands are formulated. This makes the design of our human-computer interfaces difficult and complex.

Since gestures are mostly defined by hand/arm poses and motions, a fundamental prerequisite to gesture recognition is the robust detection of hands in images. In previous work, we developed an appearance-based approach to hand detection using scale-invariant salient region features. This showed promising results, but suffered from a large number of false positives. In this paper, we focus on improving the robustness of this image-structure based approach by integrating skin color information, therefore taking advantage of the fusion of two different image cues.

The remainder of this paper is organized as follows: First, we present related work and explain why we chose the presented approach, which we describe in Section 3. Section 4 gives an overview of the data and experimental setup, and we present preliminary evaluation results on realistic data in Section 5. We conclude with a summary and an outlook on future work.

2 Related Work

A great number of different approaches to hand and limb detection using different kinds of visual cues have been proposed in recent years. A straightforward and simple approach that is often utilized (e.g. [1, 2]) is to look for skin-colored regions in the image. Although this is practicable and efficient given controlled (or known) lighting conditions, skin color detection is difficult to handle under changing illuminations, since the problem of model adaptation is not straightforward (for a recent survey on the problem, see e.g. [3]).

Another obvious drawback is that other objects having skin-like colors in the chosen representation cannot be differentiated from real skin, and therefore will yield false detections. So, given mostly unconstrained real-world scenarios, skin-color detection is not feasible to be used as a stand-alone cue for hand detection. However, since it is so simple and fast, it is attractive to be integrated as additional cue in a multi-cue detection system, because it will reliably reject false positives from other detectors that lie on non skin-colored regions.

Another widely used approach is to model an object by its shape, boundaries or general appearance, i.e based on image structure. Well-known examples are the appearance-based object detector of Viola and Jones [4] or Cootes' and Taylor's Active Appearance Models [5]. However, for strongly articulated objects – like hands – showing a large variety of shapes, this is not feasible. Describing all possible appearances of a hand as a whole would either require a very flexible model (which very likely will be too general to be still reliable) or a huge model database that would be very difficult to handle. In our research, however, we want the gestural interaction to be as unconstrained and natural as possible, which forbids reducing the variety of possible hand postures to a predefined "command alphabet".

A more promising approach to structure-based hand detection is modelling the hand as a set of characteristic parts (or image regions). If one is able to identify a set of small regions that are characteristic for hands, the input images can be searched for occurences of these regions and, if several of them being spatially close are found, it can be concluded that this image area contains a hand with high probability. We developed a system based on the salient region detector and descriptor known as SIFT (Scale Invariant Feature Transform, [6]), which uses histograms of gradients calculated around interest points extracted in Gaussian Scale Space. Note that our system does not explicitly require SIFT, any other feature extraction scheme based on local salient regions may be used. The results are satisfactory, but the approach typically yields a large number of false positives.

3 Our Approach

Since most approaches relying on a single type of feature exhibit drawbacks in specific fields or under certain circumstances, it is intuitive to think about the integration of different features. This is often referred to as sensor fusion or multi-cue integration (cf. e.g. [7]).

Provided that the different cues show substantially different characteristics and behavior, the assumption is that the strengths of one of them can compensate for the weaknesses of others, and vice versa. Consequently, this should yield a system that shows better results, reliability and robustness than any of the respective subsystems. In our scenario, it seems reasonable to utilize a skin-color detector to reduce the number of false positives produced by the structure-based region classifier. However, the answer to the question how the different features should be combined to achieve the best result is, in general, not straightforward. In this paper, we will investigate different alternatives.

In the following, we first describe the underlying parts of the classification approach,

namely structural detection using SIFT, and skin color based filtering using GMMs. Following this, the integration of both parts into a single classifier system is outlined.

3.1 Hand detection with SIFT

In our previous work we developed an approach for hand detection in realistic scenes using SIFT. These descriptions are automatically extracted by salient keypoint detection and classified via modified nearest neighbor matching. Following this, effective candidate filtering is performed by analyzing lists of keypoints spatially connected to the candidates in question.

In order to generate these adjacency lists three different approaches were investigated: circular regions of fixed size centered around the candidate; circular regions with sizes proportional to the SIFT scale of the candidate; and taking the n spatially nearest neighbors. Hand detection using local descriptors yields promising results which still, however, contain a quite high number of false positives.

3.2 Skin color classification

For an effective reduction of the number of erroneous classifications, in this paper we combine the results of SIFT-based hand detection with those provided by a skin color classification approach. For this purpose we trained Gaussian Mixture Models using a small set of training images (46 samples in full PAL-resolution) recorded within our target scenario (see section 4). The data was manually labeled with respect to skin color, and then used to train two different classifiers (using the L*a*b and normalized RG color space, respectively).

For robust detection of skin colored regions a small number of mixtures (L*a*b: 5 for skin, 16 for background; nRG: 5/2) is sufficient. Clearly, without incorporating further structural knowledge, a discrimination between hands and (parts of) non-hand regions, which might also be skin-colored, is difficult to achieve.

3.3 Data fusion

In order to integrate the abovementioned structural and color cues into a single classifier system which can be used for robust hand detection in video images, in this paper we investigated two different approaches.

A straightforward approach is to incorporate the skin color map computed using GMMs within a pre-processing step. It is then used to determine the region of interest for further processing. Alternatively (but principally identically), fusion can be performed as post-processing where the SIFT keypoints are weighted by the skin color probability.

The second integration approach we considered consists of the combination of saliency maps which are calculated separately for both information cues. Every keypoint (either SIFT or skin color related) serves as origin of a single Gaussian whose variance is dependent on the particular scale of the SIFT-keypoint, or the particular skin color probability, respectively. For their actual combination the structural and the color based saliency maps are multiplied.

4 Experimental Evaluation

For the experimental evaluation of our approach a dataset of 466 color images with PAL resolution was recorded inside our smart house over different days and under varying lighting conditions. Four persons wandering around and gesticulating (unconstrained) appeared in different distances to the cameras. The images were segmented into hand and non-hand parts manually. From this set, 93 images were randomly selected for testing whereas the remaining data was used for classifier training. Note that the GMM-training for skin color classification (section 3.2) was performed on an alternative set.

Informal experiments showed that using binary (morphologically closed) skin maps as pre- or postprocessing step is not suitable for our approach. This is because many of the structural descriptors that describe hands lie in fact outside the skin area. Also, since a skin color segmentation will never be perfect and will still exhibit holes inside skin areas, some true positives lying on skin will also be discarded. Therefore, we concentrated on the saliency map approach.

Besides the initial NN classification threshold on the SIFT keypoints, we may vary the

type of candidate filtering applied (see section 3.1), the weighting of skin- vs. keypoint saliency map prior to combination, and the final classification threshold on the combined map. In Fig. 1 we show the best results for both skin classifiers, depicted as ROC curves. The variating parameter is the classification threshold on the combined saliency map. The original ROC curve emerging from NN matching of the SIFT descriptors is also displayed for comparison.

As can be seen, the combination of our two cues greatly reduces the number of false positives, still very high true positive rates can be achieved. We marked two "working points" in the figure, with a true positive rate of 90% and 94% (the latter is the best true positive rate achieved in combination with the L*a*b skin classifier).

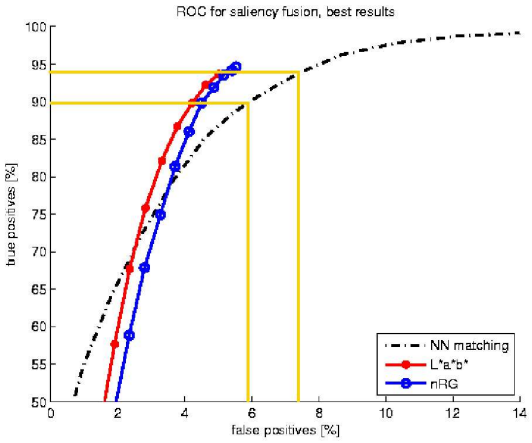
The false positive rate for 90% true positives could be reduced from 6.3% to 4.2% (a relative reduction by 33%) using the L*a*b colorspace, and to 4.5% (relative 28%) using nRG. For 94% true positives, the reduction is from 7.5% to 5.0% (33%) and to 5.4% (28%), respectively. Note that on our test data, the nRG classifier was very "optimistic" in classifying skin (i.e. it achieved a high skin recognition rate for the price of a high false positive rate) while the L*a*b* classifier was "pessimistic" (very few false positives, but discarding around 50% of true skin pixels). The fact that the final results for these classifiers do not differ dramatically clearly shows the benefit of feature fusion.

5 Discussion

In this paper, we presented an approach to hand detection using a combination of structural and color cues. Our preliminary results show that incorporating these different feature types clearly improves the quality of our hand detection system in terms of false positive rate while still achieving high true positive rates.

In future work, we will investigate other approaches to fusing the two cues. Here, we are especially interested in integrating both types of features in a single feature vector and designing a high quality, efficient classifier for these combined features. Furthermore, note that the skin detection method utilized for this paper is very simple and does not adapt to changing illumination. We expect that a more robust, adaptive skin color detector will further improve our results, and will investigate this as well.

Figure 1: Best results for saliency fusion.



References

[1] N. Hofemann et al. Recognition of deictic gestures with context. In *Proc. 26th DAGM Symposium & LNCS Vol. 3175, Springer*, pages 334–341, 2004.

[2] R. Lockton and A. W. Fitzgibbon. Real-time gesture recognition using deterministic boosting. In *Proc. British Machine Vision Conference*, pages 817–826, 2002.

[3] P. Kakumanu et al. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122, 2007.

[4] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[5] T. F. Cootes and C. J. Taylor. Statistical models of appearance for medical image analysis and computer vision. *Image Processing - Proc. of SPIE*, 4322:238–248, 2001.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60:91–110, 2004.

[7] A. Micilotta et al. Real-time upper body detection and 3D pose estimation in monoscopic images. In *Proc. European Conference on Computer Vision*, pages 139–150. LNCS 3953, Springer Verlag, 2006.