Analysis of Head-Mounted Wireless Camera Videos for Early Diagnosis of Autism

Basilio Noris[1], Karim Benmachiche[1], Julien Meynet[2], Jean-Philippe Thiran[2], and Aude G. Billard[1]

[1] Learning Algorithms and Systems Laboratory, EPFL, Switzerland
(http://lasa.epfl.ch) basilio.noris@epfl.ch
[2] Signal Processing Institute, EPFL, Switzerland

Summary. In this paper we present a computer based approach to analysis of social interaction experiments for the diagnosis of autism spectrum disorders in young children of 6-18 months of age. We apply face detection on videos from a head-mounted wireless camera to measure the time a child spends looking at people. In-Plane rotation invariant Face Detection is used to detect faces from the diverse directions of the children's head. Skin color detection is used to render the system more robust to cluttered environments and to the poor quality of the video recording.

1 Introduction

Early diagnosis of Autism Spectrum Disorders is a central topic of research in developmental psychology. Several experimental protocols for the diagnosis and understanding of developmental disorders make use of video footage analysis to measure such elements as response time, attention changes and social interaction. However, up to now the analysis of such videos is processed by manually sifting through video frames and marking timestamps and events by hand. Several applications exist that help make this marking easier but the process is still heavily time consuming.

When trying to understand the social interaction abilities of the infant, one of the most informative cues is to observe whether she looks at the people surrounding her. This can be approximated by detecting faces in the head mounted video input. Face Detection has been a central research domain in computer vision for the last decade. Since the introduction of boosted cascades of weak classifiers[1], face detection has made a leap in terms of speed and performance and has become a robust tool for a large number of applications. Further works have improved this method by adding ulterior sets of features such as tilted Haar-like features[2], Gaussian filters[3] or Local Binary Patterns[4].

In this paper we recorded videos from a wireless camera placed on the head of children playing in an unconstrained environment to gather information on the amount of time they spent looking at the faces around them. We used an in-plane rotation invariant face detection to take into account the wide range of possible orientations of the child's head. We combined it with skin color detection to limit the detection of faces to the zones of the image containing people.

2 WearCam

Recently we presented a head-mounted wireless camera that helps measure the visual attention of young children[5]. The WearCam (shown on Fig. 1) is a lightweight wireless camera mounted on the forehead of the child, thus giving a first-person point of view image that can give an approximation of the direction of attention of the child. The WearCam uses a TX45Light CCD sensor, designed to be used on miniature aircrafts, its lens has a diagonal field of view of 92°. The camera with wireless transmission electronics measures 27x27x38 mm. The sensor records an interlaced image of 640x480 pixels at 25 frames per second (fps). In interlaced videos, a single image contains two frames recorded at doube framerate, effectively yielding a 50fps video, which is essential to sustain the child's head movement. However this reduces the resolution of the image by a vertical factor of two. The battery (8.4V NiMH rechargeable) is placed on the back of the head to balance the weight of the optics on the forehead. The weight of the whole camera, battery included, is 42 grams. The camera is designed to be worn by children aged between 6 months and 2 years. The head perimeter for children in that age range varies between 35 and 48cm[1]. Adjustable straps allow the WearCam to be fit on the head of the child. Optionally the WearCam can be attached to a cap to render it easier to place on the child's head.

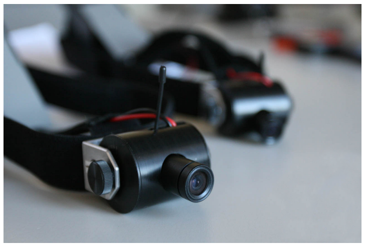

Fig. 1. The WearCam, a 42 grams head-mounted wireless camera.

3 In-plane rotation invariant Face Detection

Most applications of face detection usually deal with upright frontal and out-of-plane rotated (profile) faces. Images are normally taken from stationary or

[1] measures from the Swiss National Institute of Health

level cameras and therefore the faces present in the images are rarely subject to strong in-plane rotations. With head-mounted cameras this is not necessarily true: the angle of view of a child can vary greatly depending on her position (sitting on the floor, on a tall chair, etc.) as well as the direction she is looking at (see Fig. 2).

Some technical challenges have to be taken into account when using the WearCam. *Wireless transmission* problems can appear as distortion and blackouts for a couple of frames or as a layer of noisy interference over the whole image, which can greatly decrease the performance of appearance/shape based detection methods. *The illumination* coming through the camera can undergo severe changes very rapidly as the child turns her head towards darker or brighter parts of the environment. This forces the video analysis to take into account a wide range of luminance variations. Additionally, *the sensitivity to color* can vary greatly depending on the amount of light present in the environment, making the chrominance response of the camera span from highly saturated colors to almost grayscale images.

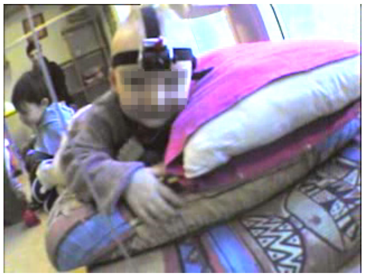

Fig. 2. A normally developing child wearing a prototype of the WearCam looking at a mirror from a slanted angle.

3.1 Boosted Cascades of Haar-Like features

The system presented here uses a set of 12 cascades of haar-like features trained on frontal faces at angles with steps of 30° of in plane rotations. Each cascade was trained from a set of 5000 rotated frontal faces and 3000 negative samples using the Gentle Adaboost algorithm, the sample size was set at 20x20 pixels. Some random rotation was added to the face samples in order to span the 30° range covered by each cascade. The cascades were trained as part of [6]. The additional set of tilted features introduced by [2] was used for all angles except the 0°, ±90° and 180° angles where only the original set of Haar-like features was used. In each case the cascade is stump-based, with 21 to 23 stages, counting between 1480 and 1785 features depending on the angles. An additional cascade for profile faces is used to improve the range of detection. The profile cascade is part of the Intel OpenCV library[2].

[2] Intel's Open Computer Vision Library, release v.1.0 http://sourceforge.net/projects/opencvlibrary/

Benchmarks

Tests on the CMU-MIT[7] benchmark were run for the multi-view face detector and compared with the default OpenCV cascades(Fig. 3). No improvement on the upright frontal faces (CMU-MIT Set I) is noticed. As expected, the detection of rotated faces (CMU-MIT Rotated) improves substantially. However, the amount of false positives increases.

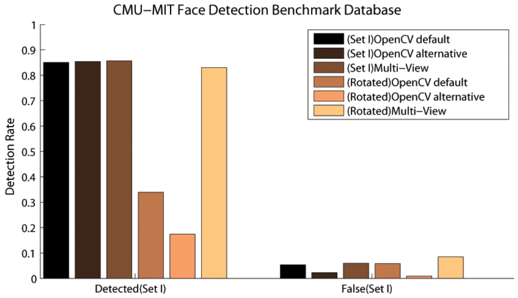

Fig. 3. CMU-MIT face database benchmark. Comparison between the *default* OpenCV face detection cascade, the *alternative* OpenCV cascade (also part of the OpenCV distribution) and our combination of rotated cascades. A candidate region is considered valid if all features (eyes, mouth, nose) are inside the region and if the width of the region is less than twice the distance between the eyes.

3.2 Pruning of false detections

A drawback of using multi-view detection with a high number of cascades is that the amount of false detections increases. In cluttered environments such as children daycare centers this becomes an important limitation. In order to overcome this problem, skin color detection is used to mask the portions of the image which are more prone to false detection (e.g. colored drawings on the walls, bookshelves, etc.). The topic of skin detection has been extensively investigated in the past[8]. However no unanimous verdict as to which method and colorspace yields the best results exists and the methods should be chosen depending on the application.

The database we used for our tests was taken from [9]. The skin dataset contains 12'250'000 samples from people of different races and ages under variating lighting conditions, while the non-skin dataset contains 25'000'000 samples. Several methods were tested (Histograms, SVM, RVM and MLP) on different colorspaces (RGB, HSV, YCbCr, SCT[10]). As most previous works[8] agree that dropping the luminance informations degrades the performance, we kept both the crominance and luminance information for all colorspaces. Kernel-PCA was used to pre-process the skin and non-skin samples using linear, gaussian and polynomial kernels (Fig. 4 shows the projection of the RGB colorspace in different kernel spaces). The performance of the best classifiers and colorspace is given in Table 1.

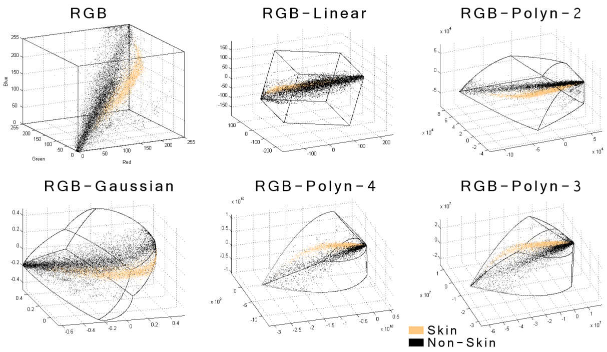

Fig. 4. Example of Kernel-PCA projections of skin and non-skin samples

The performances of the different methods selecting their respective best colorspaces are comparable. The best results were obtained using a histogram based classifier on the YCbCr colorspace. The solution presented in this paper uses a 64-bins histogram, the histogram resolution was obtained by cross-validation. Only the positive (skin) samples were used for training. For each video frame, a backprojection image of the skin histogram is computed using

$$BackProj(x, y) = hist_{skin}(y_{x,y}, cb_{x,y}, cr_{x,y})$$

where $hist_{skin}$ is the skin histogram and $y_{x,y}$, $cb_{x,y}$, $cr_{x,y}$ are the values of the pixel at position (x, y) in the image quantized to the histogram bin resolution. This yields a grayscale image that can be used as a mask for the face detection. If a face candidate does not contain a sufficient amount of skin pixels, the candidate is rejected. Fig. 5 shows the ROC curve of the skin pixels density necessary for a candidate to be considered as a valid detection. A threshold was set by elbow rule at 0.19. Due to the noisy nature of the input videos, no connectivity information on the skin color mask could be used succesfully.

4 Experimental setup

Recordings with the WearCam were made on 18 normally developing children (8 girls, 10 boys) between 2.5 and 4.5 years of age (mean 3 years and 4 months).

	SVM	RVM	MLP	HIST
RGB (poly3)	0.088	**0.059**	0.082	0.250
HSV (no-pca)	0.263	0.262	0.078	**0.058**
YCbCr (gauss)	**0.062**	0.080	0.092	0.180
YCbCr (no-pca)	0.137	0.130	0.074	**0.057**

Table 1. Classification error of the best performing methods and colorspaces for skin-detection

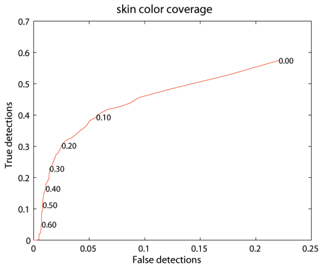

Fig. 5. Receiver Operator Characteristic (ROC) curve for skin coverage percentage of face detection regions.

The children sat around a table in pairs, supervised by two adults, and were let play with the Sony Aibo[3] robot dog. The Aibo was set in its default pre-programmed behaviour, which involves responding to petting and stroking, obeying orders given through cards held in front of its head and playing with a colored ball or bone. Both children wore the WearCam and played for a duration of 8 to 14 minutes.

The ground truth data was generated by manually tracking faces in the videos using Adobe After Effects. The manual labelling took between 45 minutes and one hour for every minute of video footage, depending on the amount of head movement of the child and the number of faces appearing in the video. A total of 74511 faces were labelled in 62597 frames of video.

4.1 Results

Table 2 shows the results of running the system on 41 minutes of the recorded videos. The detection without skin pruning correctly found 55.4% of the faces, false detection amounted to 26.3% of all the detections. The detection using skin color pruning found 51.7% of the faces, reducing the false detections to 4.3% of all detections, thus reducing significantly the amount of false alarms while decreasing only slightly the performance of the detection.

	faces detected	false detections
w/o skin pruning	55.4% ± 9.4%	26.3% ± 13.4%
skin pruning	51.7% ± 8.3%	4.3% ± 3.3%

Table 2. Results of face detection on 41 minutes frames of video

5 Discussion and Future Works

We have presented a solution for the automatic analysis of videos from a head-mounted wireless camera for the evaluation of social interaction measured as

[3] http://www.sonydigital-link.com/aibo/

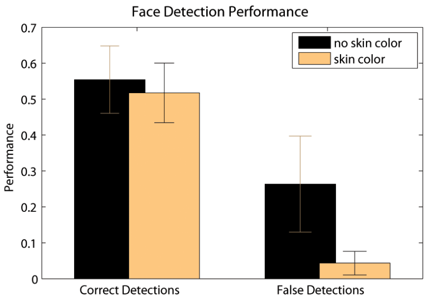

Fig. 6. Face detection performance with and without skin color detection

the proportion of time a child spends looking at faces. A rotation-invariant multi-view face detection using a boosted cascade of Haar-like classifiers was combined with histogram based skin detection to decrease false detections.

The system was able to detect more than 51% of the faces appearing in videos from free play in a cluttered environment. The skin color detection allowed to decrease the false detection rate by a factor of 6.

Compared to state of the art results the performance of our system might seem very low. However most systems run on data coming from constrained environments (e.g. uncluttered backgrounds, uniform illumination, stationary cameras) where the amount of movement or the quality of the image can be kept under check. Due to the very nature of the experimental goal (i.e. studying the behaviour of the infants in as unbiased a manner as possible) this is not possible in our case. The intense motion of the head, the wireless transmission noise and the sudden brightness and color intensity changes render the analysis of the WearCam videos challenging. The constraints the system must respect in terms of weight, dimensions and mobility do not allow the use of better quality cameras or wired transmission. Under these limitations, the results are at least promising.

One of the major issues of the system is that although mostly all the faces appearing in the video are detected fairly frequently, the detection is not continuous throughout the frames. Local spatio-temporal tracking of the detected faces should improve this and will be the main focus of future investigations.

Acknowledgements: This work was supported by the Thought in Action (TACT) project, part of the European Union NEST-Adventure Program, and by the Swiss Science Fundation within the National Center for Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2). The authors would like to thank Rainer Lienhart for his precious contribution and kind availability in providing trained cascades for face detection at different angles.

References

1. Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.
2. R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings IEEE Conf. on Image Processing*, pages 900–903, 2002.
3. Julien Meynet, Vlad Popovici, and Jean-Philippe Thiran. Face detection with boosted gaussian features. *Pattern Recogn.*, 40(8):2283–2291, 2007.
4. Guillaume Heusch, Yann Rodriguez, and Sebastien Marcel. Local binary patterns as an image preprocessing for face authentication. In *FGR '06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 9–14, Washington, DC, USA, 2006. IEEE Computer Society.
5. L. Piccardi, B. Noris, G. Schiavone, F. Keller, C. Von Hofsten, and A. G. Billard. Wearcam: A head mounted wireless camera for monitoring gaze attention and for the diagnosis of developmental disorders in young children. In *RO-MAN '07: Proceedings of the 16th International Symposium on Robot and Human Interactive Communication*, 2007.
6. R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. Mrl technical report, Intel Labs, Dec 2002.
7. Henry Rowley, Shumeet Baluja, and Takeo Kanade. Rotation invariant neural network-based face detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 1998.
8. P. Kakumanu, S. Makrogiannis, and N. Bourbakis. A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122, 2007.
9. G. Gomez, M. Sanchez, and Luis Enrique Sucar. On selecting an appropriate colour space for skin detection. In *MICAI '02: Proceedings of the Second Mexican International Conference on Artificial Intelligence*, pages 69–78, London, UK, 2002. Springer-Verlag.
10. M.W. Powell and R. Murphy. *Position estimation of microrovers using a spherical coordinate transform color segmenter.* Fort Collins, CO, Jun 1999.