Improving Face Detection with TOF Cameras

Dan Witzner Hansen*, Rasmus Larsen * and Francois Lauze†,
*Technical University of Denmark
{dwh,rl}@imm.dtu.dk
†Nordic Bioscience Imaging
francois@nordicbioscience.com

***Abstract*— A face detection method based on a boosted classifier using images from a time-of-flight sensor is presented. We show that the performance of face detection can be improved when using both depth and gray scale images and that the common use of integration of hypotheses for verification can be relaxed. Based on the detected face we employ an active contour method on depth images for full head segmentation.**

I. Introduction

For object detection and recognition to be useful in real-world applications such as robotics, surveillance, or video indexing, recognizers must have the ability to localize objects of interest in images under viewpoint changes (e.g. changes in object size or position) and must be robust to complex background clutter and preferably fast. The physical size of the objects is often restricted, but appears on multiple scales due to the relative position of camera and the object. Face detection is a particular example of such an object class. Although it has been studied for more than 30 years, developing a fast and robust face detection system that can handle the variations found in different faces, such as facial expressions, pose changes, illumination changes, complex backgrounds, and low resolutions, is still a challenging research topic. The size of the face does not vary much between subjects, but depending on the distance from the camera the apparent face size changes. This is complicated further by features emerging and disappearing with distance. Knowing the distance to the object may thus provide an important cue for face size normalization.

In this paper, we explore the use of cascade classifiers for face detection using both gray level and depth information obtained from a time-of-flight (TOF) camera. We introduce an additional stage to the classifier which uses depth information for size verification. We further employ an active contour model initialized by the face detection for segmenting the human head.

A. Previous Work

Advances in machine learning research have greatly influenced the developments in robust face detection. Neural network [13], support vector machine (SVM) [11], and boosting [15], [14], [7] are typical current choices of learning-based methods.

Current research is focused on feature extraction and appropriate structures for combining classifiers. Many types of features that have been used, ranging from simple ones such as intensity values [13], [11] and eigenspace [9] to complex ones such as wavelets [14] have been used. Face detectors based on single classifiers such as SVM [11], [12] and neural network [13] are usually slow because they process non-face and face regions equally in the input image. To deal with the problem of processing a large number of patterns, a combination of simple-to-complex classifiers has been proposed [12], [14], [5]. Viola and Jones introduce a fast object detection based on a boosted cascade of Haar-like features [14] and catalyzed a range of related papers. Lienhart extended the haar-like features to an efficient set of 45 rotated features and used discrete AdaBoost, real AdaBoost and gentle AdaBoost for face detection. More complete reviews of face detection methods can be found elsewhere [18], [16].

Several researchers have used depth cues for face tracking. Yang and Zhang [17] have applied head tracking using stereo vision. The method depends on the brightness information due to the nature of stereo imaging and, thus, it is sensitive to cluttered backgrounds or illumination conditions. Malassiotis and Strintzis [8] proposed a head tracking algorithm based on range images obtained using color coded structured light. Their work models the images using a Gaussian mixture of head and torso. A limitation of structured light is that it may be disturbing. Gokturk and Tomasi [4] propose a 3D head tracking method using correlation on large feature vectors of point-wise means and variances obtained from a time-of-flight sensor. Initial investigations of the use of time-of-flight sensors for people tracking is proposed by Bevilacqua et al. [1].

B. Time of flight technology

Time of flight sensors are a relatively new and novel development in imaging devices providing real-time gray scale and depth information from a single sensor. In this paper we use the SwissRanger (SR3000) from Mesa (www.swissranger.ch). The SwissRanger (see Fig. 1) emits sinusoidal modulated IR light with frequency f_m. Through the reflected wavefront the camera can obtain distance measurements for each pixel position by measuring the light travelling time between emission and reception. In fact, if $s(t) = \sin(2\pi f_m t)$ is the

This work is partially funded by the European network ARTTS (www.artts.eu). The ARTTS project is funded by the European Commission (contract no. IST-34107) within the Information Society Technologies (IST) priority of the 6th Framework Programme. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

emitted light, the reflected and received light measured by the sensor with a phase shift ϕ is: $r(t) = \sin(2\pi f_m t - \phi) = A\sin(2\pi f_m(t - \frac{2d}{c}))$, where A is the amplitude of the reflected light, d the distance between the target and the sensor and c is the constant of the speed of light (3×10^8 m/s). The distance can thus be calculated by $d = \frac{c\phi}{4\pi f_m}$ for each measuring point in the image (pixel). The brightness of a pixel is measured through the amplitude (A).

Compared to standard stereo vision, time-of-flight sensors offer several advantages as only a single sensor is required. No calibration between the cameras or additional image processing is needed for obtaining the depth measurement. Time of flight sensors do not require textured objects. Compared to structured light methods the time-of-flight cameras emit much less light, thus neither the image nor a person facing the camera are disturbed. The major limitations of current time-of-flight sensors are the reduced resolution (176×144 for the SwissRanger) and that depth measurements may be influenced by the reflection angle, object material and color.

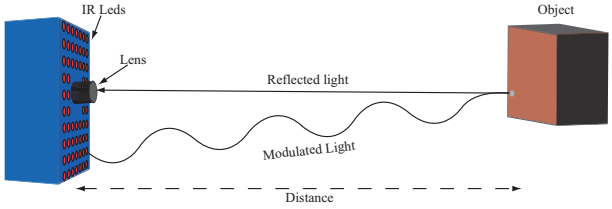

Fig. 1. Modulated light is emitted from IR LEDS on the sensor. Light is reflected on the object and captured by the sensor. The time between emission and reception and the measured amplitude is used to generate depth and intensity images.

II. Method Overview

In this paper we propose a face detection and head segmentation method based on depth and gray scale images obtained from a time-of-flight camera. Initially a set of hypothetical face regions are extracted through a boosted classifier similar to the one proposed by Viola and Jones [14], [7]. To reduce the number of false positives and to generally improve robustness, face detectors may group (integration of multiple detections) regions if they are located close. If sufficiently many hypotheses supporting a region are found the region is assumed to be correct. This approach may reject valid regions if the support is insufficient. We propose to use depth information to resolve the ambiguities without defining a support. However, if several hypothesized regions are sufficiently close, they are combined into one. Based on a detected face region, the head is segmented through an active contour model [3]. Since the depth variations within the face region are limited, we employ an active contour model that uses the depth variations in the interior and exterior of the contour to perform the segmentation.

Section III describes the extended boosted face classifier and section IV the active contour model. In section V the results of the classification and head segmentation are presented. The paper is concluded in section VI.

III. Cascaded classifier

Informative and discriminative features usually increase the detection rate and may reduce the complexity of the training methods. In a face detector that is scale and location normalized, the number of analyzed patterns is usually large (approximately 160,000 patterns for a 320×240 pixel image) because the face classifier needs to scan over the input image at every location and every scale. However, the vast majority of the analyzed patterns are non-faces. The number of regions containing faces is usually low and it thus becomes obvious that it is important to reject the majority of regions as fast as possible, while avoiding carelessly rejecting face regions.

As popularized by Viola and Jones [14], the rarity of positive examples in object detection tasks can be exploited for computational efficiency via the cascade architecture (Fig. 2). Each stage of the cascade either rejects an input region immediately as a non-object, or passes it on to the next stage for further analysis. Inputs which pass through all classifier stages are accepted as object instances. The cascade is efficient because most instances are non-objects and can be rejected by the first few stages with a minimal amount of computation. In this approach, the running time of the detector is no longer simply a function of the size of the image but also reflects the image's complexity. Blanchard and Geman [2] present a general theoretical analysis of such classifier systems.

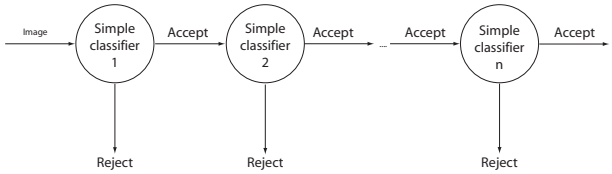

Fig. 2. A cascade of simple classifiers. Each stage either rejects a region or passes it on to the subsequent stage for further verification.

AdaBoost is used to select discriminative and significant features from a large number of features and construct the classifier. Haar-wavelet-like classifiers are used at each stage. Each simple classifier is efficiently calculated through integral images. In this paper we add a final step to the cascaded classifier where the information from the depth image and a prior model of the face sizes are used to remove invalid hypothetical regions. The prior model uses anthropomorphic averages of head sizes of a training set and compares them with the average depth within the region. The training set consists of 10 annotated images of faces captured with the SR3000.

Using similar triangles the apparent area of a fronto-parallel planar surface is $a = \frac{fA}{Z}$, where f is the focal length, A is the true area and Z is the distance from the camera to the surface (Fig. 3). The depth variations within a face are relatively small compared to the head size and, at some given scale, the face is approximately planar.

Therefore, measuring the average depth, $\bar{Z}$, within a hypothesized region and multiplying this with the apparent size

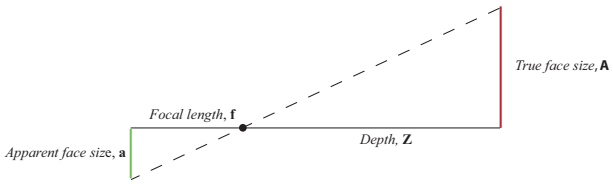

Fig. 3. Projection of a planar area (red) onto the image plane (green).

provides a measure which is proportional to the true size of the face. The focal length of the SR3000 is fixed and can thus be disregarded for classification.

The last stage of the classifier therefore classifies a face region Ω as face when

$$|a * \bar{Z}|_M \leq \tau$$

where, $|\cdot|_M$ is the Mahalanobis distance to the training set (i.e. using the mean and variance), $\bar{Z}$ is the average depth within Ω.

IV. Active Contour

Active contours are used for automatic object segmentation. The basic idea is the evolution of a curve, or curves subject to constraints from the input data. The curve should evolve until its boundary segments the object of interest. This framework has been used successfully by Kass et al. [6] to extract boundaries and edges . One potential problem with this approach is that the topology of the region to be segmented must be known in advance. An algorithm to overcome these difficulties was first introduced by Osher and Sethian [10]. They model the propagating curve as a specific level set of a higher dimensional surface. It is common practice to model this surface as a function of time. So as time progresses, the surface can change to take on the desired shape.

Several methods, such as snakes, use edge information to determine when to stop the evolution. For depth images of faces the contour boundaries give some indication of the boundary of the head, but there may be other depth discontinuities in the background as well as parts of the boundary in which the discontinuities are weak. Since the depth variations within the region are relatively limited, we examine an edge-free contour model. Rather than basing the model on an edge-stopping function, the curve evolution is determined through an energy minimization approach [3].

Let Ω be a bounded open subset of $\mathbb{R}^2$, with $\delta\Omega$ as its boundary. The image u_0 is defined by $u_0 : \Omega \rightarrow \mathbb{R}$. Consider the evolving curve $\mathcal{C}$ in Ω, as the boundary of an open subset $\omega \subseteq \Omega$ with $\mathcal{C} \equiv \partial\Omega$. The main idea is to embed the propagating curve as the zero level set of a higher dimensional function ϕ defined by: $\phi(x, y, t = 0) \pm d$ where d is the distance from (x, y) to $\partial\Omega$ at $t = 0$, and the sign is chosen to indicate inner and outer regions. Evolving the curve in the direction of its normal amounts to solving the partial differential equation [10]:

$$\frac{\partial\phi}{dt} = F|\nabla\phi|, \phi(x, y, 0) = \phi_0(x, y)$$

where the set $\{(x, y), \phi_0(x, y) = 0\}$ defines the initial contour, and F is the propagation speed. For certain forms of the speed function F, this reduces to a standard Hamilton-Jacobi equation. In this case the normal vector, n, for any point on the curve C is given by:

$$n = \nabla\phi$$

and the curvature K is obtained from the divergence of the gradient of the unit normal vector to the front:

$$K = div(\frac{\nabla\phi}{|\nabla\phi|}) = \frac{\phi_x x\phi_y^2 - 2\phi_x\phi_y\phi_x y + \phi_y y + \phi_x^2}{(\phi_x^2\phi_y^2)^{\frac{3}{2}}}$$

Introducing the minimizing energy function

$$\begin{aligned} E(\mathcal{C}, c_1, c_2) = \lambda_1 &\int_\Omega \delta(\phi(x, y))|\nabla\phi(x, y)|dxdy \\ &+ \lambda_2 \int_\Omega H(\phi(x, y))dxdy \\ &+ \lambda_3 \int_\Omega |u_0(x, y) - c_1|H(\phi(x, y)dxdy \\ &+ \lambda_4 \int_\Omega |u_0(x, y) - c_2|(1 - H(\phi(x, y))dxdy \end{aligned} \quad (1)$$

where H is the Heaviside function, δ the Dirac delta function, c_i mean values in the interior and exterior of the curve, λ_i weighing parameters. The first two terms control the curve length and area, respectively. The latter two terms measures the deviation of intensities in the interior and exterior regions. The Euler-Lagrange partial differential equation of $\phi(x, y, t)$ is given by [3]:

$$\begin{aligned} \frac{\partial\phi}{\partial t} = \delta(\phi)[\lambda_1 div(&\frac{\nabla\phi}{|\nabla\phi|} \\ &- \lambda_2 - \lambda_3(u_0 - c_1)^2 + \lambda_4(u_0 - c_2)^2)] = 0 \end{aligned}$$

The solution to this can be found using the Jacobi method for solving partial differential equations [3]. However, using the curvature K leads to a simple system that can be solved without using the Jacobi method:

$$\begin{aligned} \frac{\phi_{i,j}^{n+1} - \phi_{i,j}^n}{\Delta t} = \delta_\epsilon(\phi_{i,j}^n)[&\lambda_1 K \\ &- \lambda_3(u_{0,i,j} - c_1(\phi^n))^2 \\ &- \lambda_4(u_{0,i,j} - c_2(\phi^n))^2] \end{aligned}$$

V. Experiments and Results

In this section we describe the setup and show the results of the methods. The first part of the test evaluates the performance of the face detection the latter part shows the results of using the active contour for head segmentation.

A set of 5 sequences have been tested totalling 1089 gray scale and depth image pairs. At most one face in the range of $30 - 80$ cm from the camera is present in each frame. Each face is assumed to be near frontal.

Only the last stage differs in the two methods and thus the classifier using depth (extended classifier) is able to reduce the number of false positives, but cannot perform better on the classification rate than the standard classifier. In fact both methods have a detection rate of $88,3\%$, but where the extended classifier only has 6 false positive, the standard algorithm has 62. This classification rate is slightly lower than the one reported by Viola and Jones [14]. One reason for this is that the image data is not the same. In fact, when setting the properties of the SwissRanger to automatically adapting the emitted light, the face may become overexposed when rapidly moving towards the camera. It takes about a second for the light to be adapted correctly. When over exposed, the features of the face vanish and consequently insufficient information is available for detection.

Fig. 4. Detection results under apparent scale and head pose changes. The blue rectangles show the result of the boosted classifier which are accepted by the extended classifier using depth information and the red rectangles indicate hypothetical regions suggested by the standard boosted classifier, which are rejected in the depth validation stage of the extended classifier.

Fig. 5 show the results when using the depth for face segmentation using the parameters $[\lambda_1, \lambda_2, \lambda_3, \lambda_4] = [0.5, 0, 1, 1, 0.1]$.

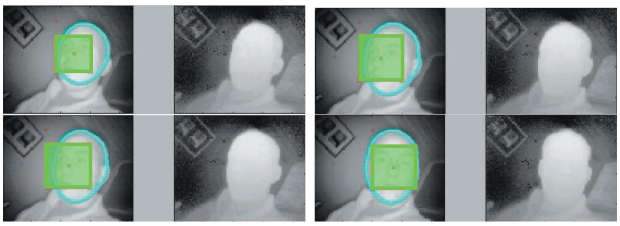

Fig. 5. The results from the head segmentation. The depth image is shown to the right of each gray scale image. The detected face is shown in green and the head segmentation is shown by the blue curve.

VI. Discussion

The time-of-flight sensor is a relatively new development in sensors for computer vision. The sensor provides both depth and intensity images in real-time without the need for tedious stereo calibration and image analysis for dense depth estimation. Even though the images are obtained in low resolution, being able to obtain both gray scale and depth images in real-time is particularly useful for many vision applications. Clearly, the combination of intensity and depth images is useful. We have presented a face detection method based on boosting and Haar features using both gray scale and depth images. Face detection on gray scale images can be done efficiently with relatively high accuracy. We show that depth information provides a reliable additional cue to face detection. When obtained from a time-of-flight sensor this added robustness is given with only negligible extra computation. We additionally suggest to use an active contour method for head segmentation using the depth image.

References

[1] A. Bevilacqua, L. Di Stefano, and P. Azzari. People tracking using a time-of-flight depth sensor. *IEEE International Conference on Video and Signal Based Surveillance, 2006.*, pages 89–89, 2006.

[2] Gilles Blanchard and Donald Geman. Hierarchical testing designs for pattern recognition. *Annals of Statistics*, 33(3):1155, 2005.

[3] T. Chan and L. Vese. Active contours without edges. *IEEE Transactions of Image Processing*, 10(2):266–277, 2001.

[4] S.B. Gokturk and C. Tomasi. 3d head tracking based on recognition and interpolation using a time-of-flight depth sensor. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2:II–II, 2004.

[5] Chang Huang, Haizhou Al, Bo Wu, and Shihong Lao. Boosting nested cascade detector for multi-view face detection. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2:415–418, 2004.

[6] M. Kass, A.P. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *Proc. Int. Conf. on Computer Vision*, pages 259–268, 1987.

[7] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *IEEE ICIP 2002*, volume 1, pages 900–903, September 2002.

[8] S. Malassiotis and M.G. Strintzis. Robust real-time 3d head pose estimation from range data. *Pattern Recognition*, 38(8):1153–1165, 2005.

[9] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[10] Stanley Osher and James A Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79:12–49, 1988.

[11] E. Osuna, R. Freund, and F. Girosit. Training support vector machines: an application to face detection. *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 130–136, 1997.

[12] S. Romdhani, P. Torr, B. Scholkopf, and A. Blake. Computationally efficient face detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2:695–700, 2001.

[13] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23, 38 1998.

[14] P. Viola and M. Jones. Robust real-time face detection. In *Proc. Int. Conf. on Computer Vision*, page II: 747, 2001.

[15] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[16] Ming-Hsuan Yang, D.J. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[17] Ruigang Yang and Zhengyou Zhang. Model-based head pose tracking with stereovision. *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 255–260, 2002.

[18] W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld. Face recognition: a literature survey. *ACM Computing Surveys*, 35(4):399–459, 2003.