INTEGRATION OF EYE DETECTION AND TRACKING IN VIDEOCONFERENCE SEQUENCES USING TEMPORAL CONSISTENCY AND GEOMETRICAL CONSTRAINTS

Gabriel M. Araujo[*]*, Eduardo A. B. Silva,*
Alexandre G. Ciancio, José F. L. de Oliveira,
Felipe M. L. Ribeiro,

Federal University of Rio de Janeiro
Cx.P. 68504, Rio de Janeiro, RJ
CEP 21945-970, Brazil
{gmatos, eduardo, ciancio, jleite, felipe.ribeiro}@lps.ufrj.br

Amir Said

Hewlett-Packard Laboratories
1501 Page Mill Road, MS 1203
Palo Alto, CA 94304, USA
amir.said@hp.com

ABSTRACT

In this work, we propose a novel approach to detect and track, in videoconference sequences, six landmarks on eyes: the four corners and the pupils. Detection is based on the *Inner Product Detector* (IPD), and tracking on the Lucas-Kanade (LK) technique. The novelty of our method consists in the integration between detection and tracking, the evaluation of the temporal consistency to decrease the false positive rates, and the use of geometrical constraints to infer the position of missing points. In our experiments, we use five high definition video sequences with four subjects, different types of background, fast movements, blurring and occlusion. The obtained results have shown that the proposed technique is capable of detecting and tracking landmarks with good reliability.

***Index Terms*—** Computer Vision, Video Tracking, Face Tracking, Eye Tracking, Object Detection.

1. INTRODUCTION

There has been an ever growing interest in augmented reality systems, specially the ones that can provide a 3D experience. These systems can either display or render multiple views of a scene or object. Then, in these cases, it is often important to know the position of the viewer's eyes relative to the display, so that the proper view can be rendered or chosen [1]. Therefore, there has been a growing need for robust eye tracking methods. In this work, we address the problem of locating and tracking, in video conference sequences, six landmarks on the eyes, namely the four corners and the pupils. Our goal is to combine features of detectors and trackers in order to deal with frame cuts, fast camera movements as well as partial and total occlusion, thereby increasing robustness of the tracking.

The proposed method presents three main contributions. The first is the integration between detection and tracking. The detection is performed using a correlation filter-based detector called *Inner Product Detector* (IPD) [2]. To track the detected points, we use the well known Lucas-Kanade optical flow tracker [3]. The proposed integration is based on the analysis of the histogram of the distances between the detected and tracked points. The second contribution further reduces the false positive rates by verifying the temporal consistency of the obtained points. Although at this step we have a high hit rate and good precision, there are missing points in several frames, leading to high false negative rates. The third contribution aims at using the geometry of the face in order to estimate the missed points.

The remaining of this paper is organized as follows. In the next Section, we present a background on the detection and tracking techniques employed. Section 3 describes how to obtain a low false positive rate by integrating detection and tracking together with an analysis of the temporal consistency. In Section 4, we devise a method to decrease the false negative rates by employing geometric constraints to estimate the position of the missed points. The experimental procedure, as well as the obtained results and their discussion, are presented in Section 5. Finally, in Section 6, we conclude and present suggestions for future work.

2. DETECTION AND TRACKING

In the next two subsections, a brief description of the employed detector and tracker is given. More detailed explanations can be obtained in [2] and [3], respectively.

2.1. Feature Detection using an Inner Product Detector

In correlation filtering, a pattern is detected by computing the cross correlation between the filter and an unknown signal [4]. The output of the filter is large when the desired pattern is input, and small otherwise. This technique has the advantage of being robust to small variations of the desired pattern. In this work, in order to detect landmarks on eyes, we use a correlation filter based detector, called *Inner Product Detector* (IPD) [2]. Suppose we define a problem involving N classes and a random variable $\mathbf{X}$, whose realization $\mathbf{x}$ belongs to a class $A_n, n \in \{1, 2, \cdots, N\}$. The objective is to obtain a detector, $\mathbf{h}_{A_n}$, for which the dot product with an unknown sample $\mathbf{x}$ is equal to one if it belongs to A_n and zero otherwise. The least squares solution for $\mathbf{h}_{A_n}$ is [2]

$$\mathbf{h}_{A_n} = \left(\sum_{i=1}^{N} p(A_i) \mathbf{R}_{A_i} \right)^{-1} p(A_n) \boldsymbol{\mu}_{A_n}, \qquad (1)$$

where $\boldsymbol{\mu}_{A_n}$ is the sample mean of the class A_n, $\mathbf{R}_{A_i}$ is the autocorrelation matrix of the samples from class A_i and $p(A_i)$ the a priori probability of a sample to belong to the class A_i. Note that, since the estimate $\mathbf{R}_{A_i}$ should be invertible, the number of samples from class A_i used in the estimate of $\mathbf{h}_{A_n}$ must be larger than the dimension of the vector $\mathbf{X}$.

Since for real world data the classes are in general not orthogonal, the inner products tend not to be strongly distributed around 0

[*]The author is also with Federal Center of Technological Education (CEFET/RJ–UnED-NI). Address: Estrada de Adrianópolis, 1.317 – CEP: 26041-271, Nova Iguaçu – RJ – Brazil

This work was partially supported by HP Brazil R&D and Brazilian agency CNPq.

and 1. In other words, the classifier in equation (1) is weak. To overcome this problem, a stronger classifier is built from it using boosting techniques [5] through a cascade of several IPD classifiers. The output of the IPD consists of several points that are candidates for a true positive detection. in the remainder of the paper, we refer to this set of points as a cloud of points. Further post-processing must be carried out in order to discard the false positives. The IPD classifiers used in this paper have been trained using the BioID database [6]. For more details, the reader is referred to [2].

3. DETECTION AND TRACKING INTEGRATION CONSIDERING TEMPORAL CONSISTENCY

One of the characteristics of the IPD detector is that, since it is essentially based on dot products, it is fast enough to be used in real time applications. In addition, as mentioned in Subsection 2.1, the output of the detector is a cloud of points. These points tend to be grouped in small clusters which are close to each other and highly correlated with the desired output [2].

In this work, we use the IPD cascade's output to feed a tracker. To track the features, we use the well known *Lucas-Kanade* algorithm (LK), an optical flow based tracker [3]. In addition, we use the consistency between the LK tracker and the IPD output in order to discard unreliable detections. In order to do so, we have to compute histograms from the cloud of points as follows:

(i) Let j be the index of the frame. Compute the vector median of the cloud of points in order to obtain a single point y_j.

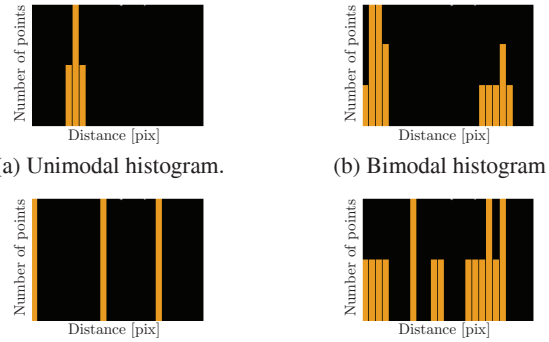
(ii) Feed the tracker with the point y_{j-1} detected in the previous frame and track it in the current one, to get the estimated point y'_j.

(iii) Let x_i ($i = 1, 2, \cdots, N$) a point in a cloud of N points and d_i the distance between the tracked point y'_j and a point x_i. Compute the histogram of the distances d_i

(iv) Analysing the histogram, we can distinguish a reliable detection from an unreliable one (see the description below about the histogram analysis). If the detection is reliable, use de vector median of the cloud y_j as output and go to step (ii). If the detection is not, do not provide an output and go to step (i).

From the analysis of the typical histograms, we can distinguish four types of behavior, which are illustrated in Figures 1(a) to 1(d). The histogram in Figure 1(a) is unimodal, which tends to correspond to a single cluster close to the desired pattern. This is an indication of a reliable detection. The histogram in Figure 1(b) has two modes. One of them will in general correspond to a cluster close to the desired point (actually, the histogram may have more than two modes, as long as they are clearly defined). This case is a candidate for a reliable detection. This is so because, as discussed later in this section, we can determine the desired cluster by analyzing the temporal behavior of the centroids of these clusters. The other two typical histograms, depicted in Figures 1(c) and 1(d), indicate unreliable detections, and the corresponding points should be rejected. The histogram in Figure 1(c), that has only a few points, appearing as several modes, is generally related with isolated noisy points. The histogram in Figure 1(d) has a "uniform" appearance (does not have any clear peak). Such behavior corresponds to a cloud with too many scattered points, that also indicates an unreliable detection.

At this point we have a temporal sequence of clusters close to the desired pattern (note that, as a result of the histogram analysis, there can be some frames with no reliable clusters). We can choose the best cluster as well as discard further unreliable clusters by analyzing the temporal evolution of these clusters. We use the strategy illustrated in Figure 2. The dots represent the points of the cloud and the crosses, the centroids of these clusters. These centroids are



(a) Unimodal histogram. (b) Bimodal histogram

(c) Histogram with few points. (d) Uniform histogram.

Fig. 1. Typical histograms obtained from an intermediate/difficult sequence for the inner corner of the right eye.

obtained by employing a Hierarchical Clustering algorithm [7] at the cascade's output. In the example in this figure, we assume that all frames up to frame $k-2$ have already been processed, and therefore all their clusters are reliable. We choose the desired cluster as the one whose centroid lies inside a disc with a radius of 4 pixels, centered in the centroid of a cluster from one of the 30 previous frames. This center is chosen as the centroid of the cluster from the closest past frame that has a reliable output. If there is no such disk in the past 30 frames, then the output of the current frame is considered as unreliable, and is thus discarded.

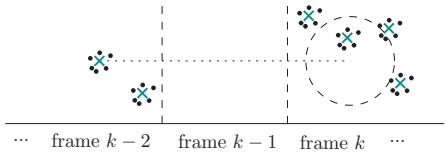


Fig. 2. Temporal evolution of the cascades output. The frames are separated by the dashed lines. The desired cluster at a frame k is selected by considering the temporal evolution of the cluster at previous frames. Note that the closest reliable cluster is at the frame $k-2$ and there is no reliable output at frame $k-1$.

The centroids of the clusters that remain reliable after the histogram and temporal consistency evaluation have a high probability of corresponding to the desired features. Therefore, we have a detector output with a low false positive rate. However, many frames are marked as having unreliable outputs, which gives rise to a high false negative rate (see the graphs on the left in Figure 5). To overcome this problem we have devised a method whereby, whenever there is no reliable detection, we determine the location of a likely detection based on the geometry of the face. This is described in the next section.

4. GEOMETRIC CONSISTENCY

As mentioned previously, the feature points of the eyes that we want to track are the left and right outer corners, left and right inner corners and the two pupils. Considering that real world faces are far enough from the camera, it is reasonable to suppose that such feature points lie on a plane. Then, the correspondence of eyes' feature points in two different frames can be described by a 2D homography $\mathbf{H}$ [8].

To obtain this homography $\mathbf{H}$, we assume that, between frames, the face is translated by $\mathbf{t} = (t_x, t_y)$ and is rotated by θ around an axis parallel to the camera's principal axis. In addition, we assume also a rotation around an axis orthogonal to the 3D scene's horizontal plane, that can be modeled as a scaling s along the camera plane's horizontal direction.

Using this motion model, the desired transformation has four degrees of freedom. As each point correspondence between frames puts two restrictions on the homography $\mathbf{H}$, we need two correspondences to determine $\mathbf{H}$. Since we can assume that the four corners of the eyes comprise a rigid body, we can use these points as references to obtain the desired homography. Therefore, if at least two eye corners have been reliably detected in the current frame, and there is a previous frame in which all eye corners have been reliably detected, the missed points from the current frame can be estimated.

Figure 3 depicts the $\mathbf{H}$ parameters $\mathbf{t}$, θ and s, where $\mathbf{x}_i$ is a point in a previous frame and $\mathbf{x}'_i$ is its correspondence in the current one.

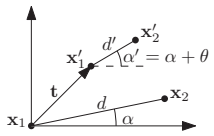


Fig. 3. Parameters necessary to compute the homography $\mathbf{H}$ between eye corners in two frames. Note that putting the reference on $\mathbf{x}_1$, we have: $\mathbf{t} = \mathbf{x}'_1$, $\theta = \alpha' - \alpha$ and $s = d'/d$.

After obtaining the parameters in Figure 3, a missed corner can be estimated using the expression

$$\mathbf{x}'_i = \mathbf{H}_r(\theta)\mathbf{H}_r(\alpha)\mathbf{H}_s\mathbf{H}_r(-\alpha)\mathbf{x}_i + \mathbf{t}, \tag{2}$$

where

$$\mathbf{H}_r(\gamma) = \begin{bmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{bmatrix}, \qquad \mathbf{H}_s = \begin{bmatrix} s & 0 \\ 0 & 1 \end{bmatrix}. \tag{3}$$

When there are more than two correspondences, each pair is used to compute a different homography. The missed point is the average of the points obtained employing each possible homography.

Since the pupils can move relatively to the eyes' corners, we have to use a different geometrical model for them. It is based on the reasonable supposition that the distance between the pupils is constant, and also that the line connecting them remains parallel to the lines connecting the two inner or the two outer corners of the eyes. We can estimate the position of a missing pupil in the current frame provided that we know: (i) the position of the eyes' corners in both frames; (ii) the position of the pupils in the previous frame, and (iii) the position of one pupil in the current frame. This model is illustrated in Figure 4. In this illustration, the right pupil is represented by a cross and the left pupil by a square. The coordinates (δ_x, δ_y) of the pupil on the right eye relative to the outer right eye corner are the same as the coordinates of the left pupil relative to the left inner eye corner. These coordinates can be used to determine the missing pupil position.

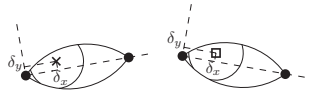


Fig. 4. Geometric model for the location of the pupils.

5. RESULTS AND DISCUSSION

In this section, we present the used database, the experimental procedure and the obtained results. In Subsection 5.1, we describe the database. The accuracy measure used to assess the performance of our method is described in Subsection 5.2. Finally, in Section 5.3, we present and discuss the obtained results.

5.1. Used database

In our experiments, we have used five high definition (1080p) videoconference sequences with 300 frames each. The sequences have a moderate degree of compression artifacts. In these sequences, we have four subjects with different skin colors, different types of background, movement and face occlusion. The sequence we refer to as "easy" has little movement of the subject and no occlusion. The one referred to as "intermediate/difficult" has blur, a subject with a moderate amount of movement and no occlusion. The other three sequences are considered difficult since they have blur, subjects with fast movements, and partial or total occlusion of the face by one of the hands. We manually annotated 13 fiducial points on the faces (that include the eye's corners and pupils) in all 300 frames. The sequences, as well as the manual annotations, are available at [9]. As this work is the first one with this database, a comparison with previous work could not be performed yet.

Note that, since we are dealing with videoconference sequences, the pre-processing scheme used by [2] is capable of dealing with illumination and scale changes. This is so because it normalizes the face size as detected by the Viola-Jones algorithm [10], and employs illumination normalization [11].

5.2. Evaluation

To assess the performance of our method, we employ an accuracy measure based on the one proposed in [12]. It is the displacement between the reference and the automatic label normalized by the correct inter-ocular distance, that is,

$$e = \frac{\|\mathbf{l}_a - \mathbf{l}_m\|}{d_{\text{eyes}}}, \tag{4}$$

where l_a is the coordinate of the automatic label given by the proposed method, l_m is the coordinate of the manually annotated point (ground truth), and d_{eyes} is the inter-ocular distance obtained from the ground truth.

For each landmark of each sequence, we computed the hit rate against the percentage of the inter-ocular distance that is considered as a correct detection. We plotted the hit rate considering: (i) only the cases for which the algorithm provides an output, taking into account only FP (False Positive) errors, and (ii) all cases, even the ones which the algorithm outputs no points, taking into account both FP and FN (False Negative) errors.

5.3. Results and Discussion

Due to space restrictions, we present in this paper only the results for one intermediate/difficult sequence. Also, since the detection results tend to be equivalent for the left and right eyes, we show the results of only three points. The remaining curves (all points of all sequences) can be seen in [9].

In Figure 5, the plots on the left were obtained without using the geometrical consistency to correct the missing points, whereas the plots on the right were obtained using the geometric consistency. The plots on the left show that the use of histogram analysis and

temporal consistency was capable of providing a very low false positive rate, but at the expense of a high false negative rate. By comparing them with the plots on the right, we can see that, by using the geometric consistency one could significantly decrease the false negative rate, at the expense of a small increase in false positive rate. These results were consistent for all sequences.

However, the analysis of the results for the difficult sequences [9] shows that, although the use of the histogram analysis and temporal consistency are able to provide a reasonably low false positive rate, the use of geometric consistency is not as effective as in the cases of easy and intermediate/difficult sequences. This happens because for such sequences there is a large number of frames where no reliable points are detected. In addition, the false positive rates are not as low as in the easier sequence. These points in errors lead to wrong geometrical models, resulting in wrong points being put in place of missing ones, thus increasing the false positive rates. However, one must note that in many practical cases the sequences used will be more like the easy and the intermediate/difficult sequences. In addition, most known methods tend to fail in the case of sequences such as our hard sequences (with blur, fast movement and occlusion by one of the hands).

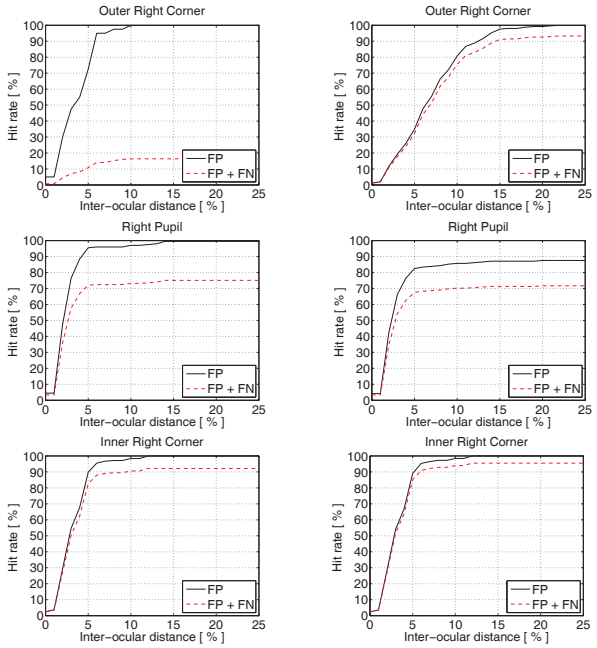


Fig. 5. Results for an intermediate/difficult sequence. Left: only detection/tracking integration and temporal consistency. Right: same as on the left, but with added geometric consistency.

6. CONCLUSIONS

In this paper, we describe a novel method to robustly detect and track six landmarks on eyes. The proposed method has two main contributions. The first one is the integration between detection (in this paper performed by an *Inner Product Detector*) and tracking (using the Lucas-Kanade algorithm). Through this integration, based on the analysis of the histogram of the distance between tracking and detection, as well as temporal consistency, we can find candidate points with low probability of false positives and reasonable accuracy. However, this is obtained at the expense of a high false negative rate. The other main contribution of this paper aims at reducing this false negative rate by employing geometrical consistency between frames. This geometrical consistency is carried out by using the face geometry to impose localization constraints and recover some missed points. Note that these results are achieved with reasonably low computational complexity. The proposed detectors have been implemented in C++ using the OpenCV library, and can run in real time on a fast PC.

The results show that the proposed method performs well for easy and intermediate/difficult sequences. However, it needs to be improved to deal with hard sequences, where we have a great deal of fast movement, blur and occlusions. It is important to point out that the integration between detection and tracking developed in this paper can be used with several tracking methods currently available. In addition, it can also be used with other types of detectors, as long as they can be adapted to output a cloud of points instead of a single point. There is a work in progress where detectors more robust than the IPD are being investigated, as well as better performing trackers, such as particle filtering [13] or the one in [14].

7. REFERENCES

[1] Cha Zhang, Zhaozheng Yin, and D. Florencio, "Improving depth perception with motion parallax and its application in teleconferencing," in *Multimedia Signal Processing, 2009. MMSP '09. IEEE International Workshop on*, oct. 2009, pp. 1–6.

[2] Gabriel Araujo, Waldir Sabino, Eduardo da Silva, and Siome Goldenstein, "Facial landmarks detection based on correlation filters," in *International Telecommunications Symposium (ITS)*, 2010.

[3] Bruce D. Lucas and Takeo Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2*, San Francisco, CA, USA, 1981, pp. 674–679, Morgan Kaufmann Publishers Inc.

[4] B. V. K. Vijaya Kumar, Abhijit Mahalanobis, and Richard D. Juday, *Correlation Pattern Recognition*, Cambridge University Press, New York, NY, USA, 2005.

[5] Robert E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, June 1990.

[6] BioID Technology Research, "The bioid face database," http://www.bioid.com, 2001.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[8] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, second edition, 2004.

[9] Eye Tracking Research, "Robust eye tracking in high definition videoconference sequences," http://www.lps.ufrj.br/∼biometria/etr/, 2012.

[10] Paul A. Viola and Michael J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[11] Xiaoyang Tan and Bill Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," in *Analysis and Modelling of Faces and Gestures*. oct 2007, vol. 4778 of *LNCS*, pp. 168–182, Springer.

[12] Oliver Jesorsky, Klaus J. Kirchberg, and Robert Frischholz, "Robust Face Detection Using the Hausdorff Distance," in *AVBPA '01: Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication*, London, UK, 2001, pp. 90–95, Springer-Verlag.

[13] Emilio Maggio and Andrea Cavallaro, *Video tracking: theory and practice*, Wiley, 2011.

[14] Z Kalal, J Matas, and K Mikolajczyk, "Online learning of robust object detectors during unstable tracking," *On-line Learning for Computer Vision Workshop*, 2009.