

Tracking Small Hand Movements in Interview Situations

Enrica Dente^{‡*}, Jeffrey Ng^{‡*}, Aldert Vrij[†], Samantha Mann[†], Anthony Bull[‡] & Anil Bharath[‡]

[‡]Imperial College London & [†]University of Portsmouth

Abstract

In this paper, we motivate ongoing work into developing methods for the automated tracking of small hand movements in interview situations to aid non-verbal behaviour analysis in the detection of deception. Existing techniques for detecting and tracking hand motion are reviewed to place current and future technical work into context. We present a modification to the popular colour predicate approach to skin detection based on Bayesian posterior probability maps and Parzen colour space probability density estimates. We demonstrate the application of a complex wavelet decomposition to identify changes in finger position. Although our existing hand tracking algorithm currently relies on posterior probability map thresholding, morphological operations and distance heuristics, we suggest the role of kinematic models of upper body, limb and finger motion for future work.

1 Introduction

Physiological, non-verbal and verbal approaches have been used in order to discriminate between “deception” and “truthfulness” [18]. Recently, non-verbal methods of detecting deception have appeared to be promising, mainly because it has been found that people who are trained at recognising visual cues of deception are better at discriminating liars from truth tellers than other non trained observers [20]. This has led to the hypothesis that the integration of improved human analysis with automated techniques for detecting and tracking visual cues of deception can help create more accurate means for detecting human deception [9].

In the analysis of non-verbal cues of deception, an individual’s baseline behaviour, such as the frequency of occurrence of eye blinks, is quantified. Any deviation from this baseline behaviour, detected during the interview, is also measured. These deviations are then analysed after the interview to determine what has caused them [1]. However, when dealing with large amounts of data, which have been simultaneously acquired from multiple

sources, it can be argued that human observation of non-verbal cues is subjective and limited. Therefore, promising results are more likely to be achieved by applying objective and systematic deception detection methods [20] [9].

In support of this idea, the calculation of the frequency of occurrence of hand and finger movements has proved to significantly increase the accuracy of deception detection in experiments carried out with professional lie catchers, compared to subjective global impressions of the body movements of the subjects interviewed [20]. In these experiments it has also been found that suspects in high stake situations do not show nervous behaviour, as most professional lie catchers believe, but rather a decrease in the frequency of hand and finger movements when they lied, due to either an increase in cognitive load [8] or to controlled behaviour [7].

Encouraging results have been reported for automated techniques, designed to detect and track visual cues of deception, which allow quantification of “micro-expressions” used in automated face analysis [3] and quantification of the thermal patterns around subjects’ faces [15]. However, people who are physiologically aroused are likely to be detected by these systems, and hence accused of lying, while, at the same time, experienced liars might not be aroused at all [19]. Although these techniques can detect “psychological states” such as fear, their reliability in discriminating between “deception” and “truthfulness” has yet to be proved.

Although a completely automated solution for non-verbal behaviour analysis in deception detection is an appealing prospect [20], there are issues over the validation of systems which stem from the great difficulty in designing rigorous controlled experiments in which one can unambiguously define a ground truth. On the contrary, the integration of more objective and systematic human observation with the adoption of automated tools for behaviour analysis in specific areas such as tracking finger movements seems to be the way forward. This paper presents a first and innovative attempt in this direction.

*This work was partly funded by the UK Research Council Grant GR/R87642/02.

2 Review of Techniques

Most computer vision approaches to human gesture tracking make use of skin detection and feature extraction techniques. Various methods have been successfully used for automated hand detection. These methods make use of two main operations. Firstly, a segmentation process is applied which aims to separate, in a sequence of images, regions containing skin tone from the background, using an accurate representation model of skin and non skin tone for classification purposes. Secondly, a model of hand motion between frames needs to be deployed for tracking the hands.

The segmentation is usually accomplished in several stages: skin distribution modelling, which includes a training process, usually supervised, followed by several morphological operations for noise removal until no further segmentation is required to create an accurate representation model of skin colour. State-of-the-art skin detection approaches do not usually include a justification of their colour space choice. This is probably because there is no colour space that can fully remove the changing lighting conditions, shadows and complex background containing surfaces and objects with skin-like colours, factors which can affect the accuracy of skin colour detection [13]. However, a colour space more suited to discriminate skin from non skin colour values can make it easier to build an accurate classifier.

Kjeldsen [12] proposed a “colour predicate” approach to skin segmentation, which relies on skin colour and non-skin colour pixel examples being used to increment and decrement histogram bins directly. This simple non-parametric approach, consists of estimating skin colour distribution from training data without deriving and updating an explicit model of the skin colour. Skin colour distribution is estimated by assigning a probability value to each point of a discretized colour space and by using a normalised colour lookup table as a “template” for finding skin tones. Additional finer segmentation has to be performed in order to distinguish the hands from the face.

Due to its simplicity, the colour predicate algorithm has been mostly used in automated deception detection, alongside additional morphological operators for detecting skin colour and for removing misclassified background pixels. However, the identification of skin colour using this simple approach is problematic because skin colour varies with factors such as lighting and view direction of the camera [13]. Also, for accurate results the choice of an appropriate colour space and of adequate decision rules need to be done empirically.

Alternative and, in some cases, more accurate methods for detecting skin tone regions make use of mixtures of Gaussians.

In order to detect hands across frames, systems need to be robust, i.e invariant to changes in lighting, shadows and occlusion [17]. Histogram-based approaches to skin colour representation usually miss a large number of colour pixel values if the number of colour samples is not representative. They suffer from difficulty in updating the representation efficiently across frames, storage problems and the “curse of dimensionality”, particularly if extra features, beyond colour, are included. Some of these problems can be addressed by using a functional form for the density model with statistical properties capable of providing an accurate representation of the density of the data.

The most promising parametric approaches model skin colour as components of mixtures of Gaussians. In these approaches the foreground, and also the background, are modelled by a joint probability density function. The contribution of each Gaussian is determined by a scalar weight, a mean vector and a diagonal covariance matrix. The model parameters (i.e means and covariance) for the Gaussians are estimated from the training data using a maximum likelihood or Bayesian inference approach. The parameters of each component of the model are updated online. This operation can rely not only on colour but also on position and motion information. The final skin probability is then computed from these Gaussian Probability Density Functions (PDF’s).

Non-adaptive methods of modelling the background need to be manually initialised on a clear frame. In order to address this problem, Stauffer and Grimson [17] model the distribution of colour values of each pixel as a mixture of Gaussians, and use adaptive background subtraction based on the observation frequency and variance of each of the Gaussians in the mixture. This approach is not suited to detecting small hand movements in static interview situations because it is based on the assumption that the foreground is moving. On the contrary, a spatial-based approach such as *Pfinder* is preferable. *Pfinder* uses a single Gaussian model per pixel for modelling the static background and builds a multi-blob statistical model of the user based on the prediction of the motion of blobs using a Kalman filter [21].

The main advantages of using mixtures of Gaussians for skin modelling are that they do not require a large storage space like the colour predicate and they provide a more compact skin model representation

and the ability to interpolate or generalize the training data. However, most of these systems require accurate initialization and assume the number of components to be known in advance [21].

Some approaches to hand detection focus on estimates of change between frames. The most common of these is image or frame differencing, a method for separating moving objects from a static background. It consists of extracting the pixels between two or three consecutive frames with a difference greater than a given threshold to detect moving objects. The moving objects are clustered into motion regions using connected components labelling. However, this technique is sensitive to noise including lighting changes occurring from frame to frame. Also, it cannot detect all the area of the objects present in a scene unless the objects undergo significant motion across frames [14]. Therefore, it is likely to be unsuitable for detecting small hand movements in interview situations.

Motion segmentation based on optical flow can also be used to detect moving regions in a frame sequence. The main advantage of using this approach is that the data acquired may be useful for the extraction of articulated objects which can be used in higher level processing. However, differential-based methods are sensitive to noise while correlation-based techniques are less sensitive to noise but require the selection of a significant number of parameters in an empirical way. Finally, most optical flow methods are computationally complex [2].

More recently, eigenspace approaches on contour and appearance feature spaces requiring examples of hand gestures, have been applied to sign language recognition and behavioral state identification. Bowden’s contour feature spaces [5] are based on the Point Distribution Model principle [6], according to which the shape and deformation of an object can be described statistically. Based on this principle, Bowden formulates the shape of a hand as a vector representing a set of points specifying the path of the contour of the hand for each model class. This shape is then learnt through statistical analysis, based on the assumption that the training set forms a cluster which is hyper-elliptical in shape. Using this assumption, Principle Component Analysis (PCA) is used to extract the centre and the bounds of the ellipse along each of its major axes by analysing the eigenvectors and values of the covariance matrix. By projecting the training set down into the linear subspace as derived from PCA, the dimensionality and computational complexity of the non-linear analysis is reduced significantly and, as a result, facilitates statistical and probabilistic analysis of the

training set. Lu used manually created sub-space training samples [13] for one and two hands to enhance the accuracy of hand detection. However, it is likely that this approach is not suited to detecting small hand and finger movements, because it is hard to segment the finger movements from the hand space target.

In order to track the hands, the most commonly used model of hand motion consists of measuring the correspondence of the centroid of the hand blobs across frames [13]. However, the self-occlusion between hands and face remains an unsolved problem when using colour information only. A resolution of this problem requires that some prior rules or knowledge about the human form and its motion be used to augment the colour information. Such knowledge is likely not only to aid tracking in the presence of such occlusions, but also to make the behavioural classification of subtle hand movements easier.

For the first stage of the segmentation, we have found that an accurate identification of skin tone pixels can be obtained by a simple Bayesian approach, which we explain in more detailed in the next section. In order to capture small hand movements, initial experiments on frame sequences of interview situations showed that complex wavelets are a promising technique for characterising finger position.

3 Skin Detection Using Simple Bayesian Posterior Maps

Noting that skin hue and saturation values are relatively stable indicators of skin colour regardless of the luminance range, and that the choice of alternative colour spaces such as normalised RGB does not make a significant difference, we use a simple Bayesian posterior map to identify the most probable locations of skin in each frame. The current system necessitates a training process, whereby example hue and saturation values are extracted by a human operator from hand and non-hand image regions. The removal of luminance provides the benefit of the reduced dimensionality of the example pixels used. We compute a joint conditional density function in hue and saturation for hand and non-hand classes using Parzen density estimation with isotropic Gaussian kernels for both feature dimensions.

$$p(H, S | Hands, \{H_n^h, S_n^h\}_{n=1..N_H}) = \frac{1}{K_H} \sum_{n=1}^{N_H} e^{-\frac{(H-H_n^h)^2 + (S-S_n^h)^2}{2\sigma_0^2}} \quad (1)$$

where H and S represent the hue and saturation variables, H_n and S_n represent the observed values amongst the N_H hands training set pixels, and σ_0

specifies the width of the kernel used in smoothing. K_H is set to normalise the density function over both variables. Similarly, from the N_N non-hands pixels:

$$p(H, S | NonHands, \{H_n^{nh}, S_n^{nh}\}_{n=1..N_N}) = \frac{1}{K_N} \sum_{n=1}^{N_N} e^{-\frac{(H-H_n^{nh})^2 + (S-S_n^{nh})^2}{2\sigma_0^2}} \quad (2)$$

where the symbols have the corresponding association with those of Equation (1), but for the non-hand pixels.

Once the results of the training are obtained, we may construct the posterior probability map for skin location by

$$p(Hands | H, S) = \frac{P(H, S | Hands)p(Hands)}{p(H, S)} \quad (3)$$

where we assume that $p(H, S | Hands)$ is approximated by the left hand side of Equation (1) and $p(H, S | NonHands)$ is approximated by the left hand side of Equation (2). This latter quantity is required to evaluate the denominator of Equation (3), since

$$p(H, S) = p(H, S | Hands)p(Hands) + p(H, S | NonHands)p(NonHands) \quad (4)$$

3.1 Hand Tracking

Presently, hand tracking is performed by finding the centroid correspondence of each object across frames from the binary image produced by a thresholded posterior probability map. Connected components analysis is used to label the blobs. Starting from a user identified hand in the first frame being tracked, the algorithm switches from a one-hand to two-hand state by using distance and area heuristics. Bounding boxes from one or two hands are then used to direct the attention of an orientation mapping algorithm based on a complex wavelet transform.

4 Complex Wavelets for Characterising Finger Position

In order to capture and represent small-finger movements, we use a rotationally-steerable [10] complex wavelet [11] decomposition [4]. Complex wavelets provide a stable representation of the orientation of local image structure in a manner which is relatively invariant to local phase. The wavelets are similar to two dimensional complex Gabor wavelets, but are designed with certain reconstruction properties [4]. The decomposition is multi-scale, but we use just the first two scales in order to capture finger position. Since the decomposition relies on a filter bank of complex wavelets, each tuned to

one of K orientations equally spaced in $[0, 2\pi)$, a vector indicating the direction of local image structure may be estimated by a weighted vector summation operation. An illustration of $K/2$ of the real and imaginary impulse responses of these wavelets, at scale 1, for $K = 8$ is shown in Figure (1).

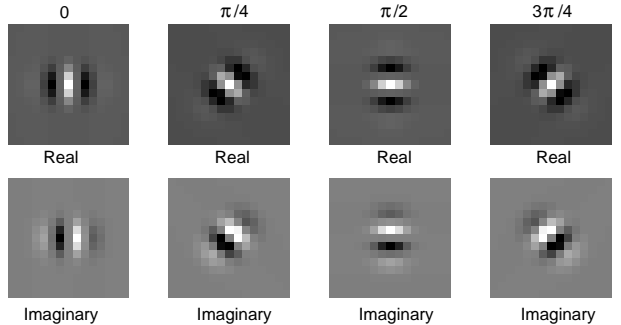


Figure 1: Illustration of the real and imaginary parts of the wavelet kernels used. They are employed in a multiscale framework, so that the effective impulse responses grow at coarser scales of analysis.

The vectors of this summation are the fixed unit vectors defining the axis of symmetry/antisymmetry of each oriented complex wavelet pair, and the weights at each position in space are the scalar fields comprised of the magnitudes of the complex outputs of the wavelet filters at each point in space. The implementation of this vector summation is through complex algebra, and the resulting vector field appears as a complex field in which local orientation is encoded by complex number phase.

Briefly, if $f_k^\ell(m, n)$ represents the set of complex wavelet coefficients produced by the decomposition, where k indexes the direction of the wavelet sub-band sensitivity, and ℓ represents the scale of the decomposition, then for each of L scales, at every position, we define the orientation field, $\mathbf{O}^{(\ell)}(m, n)$ by,

$$\mathbf{O}^{(\ell)}(m, n) = \frac{\sum_{k=0}^{K/2-1} |f_k^{(\ell)}(m, n)| e^{j2\phi_k}}{p + \left(\sum_{k=0}^{K/2-1} |f_k^{(\ell)}(m, n)|^2 \right)^{\frac{1}{2}}} \quad (5)$$

where the summation is taken over only $K/2$ directions that reflects the observation that only $K/2$ unique filters are needed in practice, the other directions being provided by the complex conjugates of the unique $K/2$ filter outputs. The denominator of Equation (5) is a “divisive normalisation” term, which reduces the effects of contrast variation. The small constant p conditions the operation, avoiding divisions by zero. The “dominant” orientation map is overlaid on the image in Figure (2), providing an indication of how well it captures hand pose.

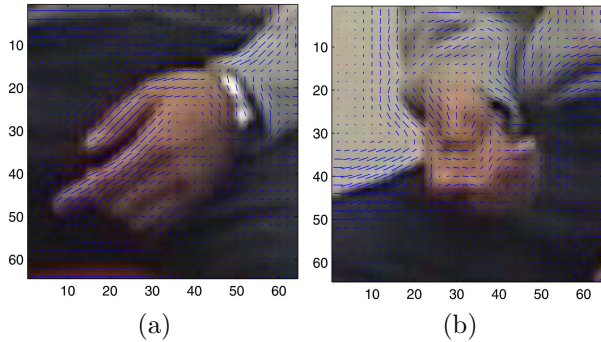


Figure 2: Two examples of orientation overlaid on hand images. Although hand pose cannot be captured entirely by this single-camera approach, there is substantial salient information contained in the orientation field, displayed by an overlay of thin bars.

4.1 Orientation Histogram

Using the posterior skin map, combined with the blobs identified during tracking, one can produce a histogram of orientations for each hand in each frame. This is done by probabilistic weighting of the orientation map, windowed by the bounding box of each blob, with the hue and saturation posterior probability map for skin and the magnitude of the orientation dominance field. This is simply implemented by a modified histogram function that contributes a weighting of the product of the orientation field magnitude and posterior skin tone map to the appropriate histogram bin (rather than a simple count of unity). The result of this, smoothed by a 7 point Gaussian kernel of variance 1, and normalised to unity integral, is illustrated in Figure (3). Note that the density functions are quite distinct for these two cases, despite the fact that the hands have remained in much the same position in the image frame. Thus, finger movement is more likely to be detected through the orientation density function rather than, say, centroid changes in blobs.

One may question whether the histogram is, in fact stable over several frames where there is a distinct lack of finger movement. We illustrate a simple experiment in Figure (4), in which 10 frames of movement are compared with a starting frame. One can see that the fluctuation of the histogram is minimal. However, a proper metric must still be determined for identifying when pairs of orientation histograms can be associated with relative finger motion.

5 Conclusions & Further Work

We have presented early work on tracking and characterising hand motion to aid non-verbal behaviour analysis in the detection of deception. With sufficient user intervention, the approach described here provides a platform from which to incrementally add sophistication. The proposed

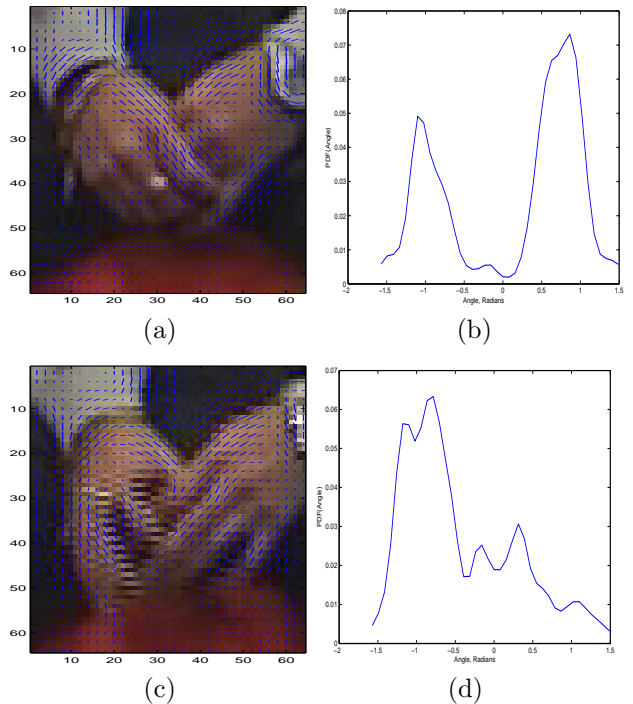


Figure 3: (a) Extracted region containing hands and superimposed orientation map. (b) weighted histogram of orientations from image on the left. (c) Extracted region containing hands just in the process of separation. (d) The weighted histogram of orientations contained in the image on the left.

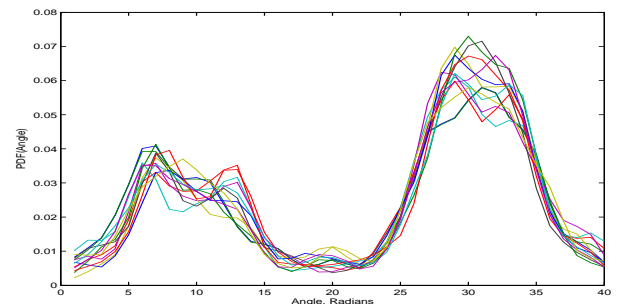


Figure 4: Illustration of the reproducibility of orientation PDF's in the absence of finger motion.

refinements are, firstly, the use of mixture models for both background rejection and for foreground (hand) tracking. Further work is also clearly needed to establish the sensitivity and specificity of the orientation histogram method to small finger movements within the wider context of hand movements. Tracking could be improved, and the need for training reduced, by equipping the system with prior knowledge on human shape and form.

One of the areas, therefore, to be explored is the incorporation of kinematic and anthropometric models of human motion and pose. Descriptions of limb dimensions and connectivity, of joint motion and stiffness characteristics are aspects under consideration.

Although the wavelet decompositions used here have been implemented in software, there are hardware versions in development which run in near-real-time. Thus, the computationally intensive overcomplete wavelet transform can be significantly “pipelined”, so that it represents only a memory/bandwidth overhead, particularly if regions of interest of, say 128x128 square pixels are used for capturing the hands. To begin with, the main application of this work is to address the reproducibility issues intrinsic to human-based markup of video sequences used in experimental psychology. In the longer term, we wish to infer more subtle aspects of subject state from hand movements [16]. In order to create better defined cues, decision models and semi-automated tools which can aid deception detection, our approach is to integrate a number of techniques, models and approaches, often drawn from different disciplines.

References

- [1] S. H. Adams. *Communication under stress: Indicators of Veracity and deception in written narratives*. PhD thesis, Virginia Polytechnic Institute and State University, 2002.
- [2] J.L. Barron, D.J. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12(1):43–77, 1994.
- [3] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In *Proc., IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2005.
- [4] A. A. Bharath and J. Ng. A complex steerable wavelet construction and its application to image denoising. *IEEE Trans. on Image Processing*, page In Press, 2005.
- [5] R. Bowden and M. Sarhadi. A non-linear model of shape and motion for tracking finger spelt american sign language. *Image and Vision Computing*, 20:597–607, 2002.
- [6] T. F. Cootes, C.J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61:38–59, 1995.
- [7] B.M. DePaulo and S.E. Kinrkendol. The motivational impairment effect in the communication of deception. *Credibility Assessment*, 22:51–70, 1989.
- [8] P. Ekman, W. V. Friesen, and K.R. Scherer. Body movement and voice pitch in deceptive interaction. *Semiotica*, 16:23–27.
- [9] J. Burgoon et al. An approach for intent identification by building on deception detection. In *Proc. of the 38th Annual Hawaii Int. Conf. on Detection of Deception: Collaboration Systems and Technology*, 2005.
- [10] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.
- [11] N. Kingsbury. Image processing with complex wavelets. *Philosophical Transactions Of The Royal Society Of London – Series A*, 357(1760):2543–2560, September 1999.
- [12] R. Kjeldsen and J. Kender. Finding skin in color images. In *Second International Conference on Automatic Face and Gesture Recognition*, 1996.
- [13] S. Lu, G. Tsechpenakis, and D. N. Metaxas. Blob analysis of the head and hands: A method for deception detection. In *38th Hawaii International Conference on System Sciences*, 2005.
- [14] J. Martin and J. L. Crowley. An appearance-based approach to gesture-recognition. *Lecture Notes in Computer Science*, 1311:340, 1997.
- [15] I. Pavlidis. Lie detection using thermal imaging. In D. P. Burleigh, K. E. Cramer, and G. R. Peacock, editors, *Proceedings of the SPIE*, volume 26, pages 270–279, 2004.
- [16] A. Psarrou, S. Gong, and M. Walter. Recognition of human gestures and behaviour. *Image and Vision Computing*, 20(5-6):349–358, 2002.
- [17] C. Stauffer and W. E. L.Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22:747–757, 2000.
- [18] A. Vrij. *Detecting Lies and Deceit: The Psychology of Lying and the Implications for Professional Practice*. John Wiley & Sons, Inc, 2000.
- [19] A. Vrij. Guidelines to catch a liar. *Deception detection in forensic contexts, Cambridge, England: Cambridge University Press*, pages 287–314, 2004.
- [20] A. Vrij and S. Mann. Detecting deception: The benefit of looking at a combination of behavioral, auditory and speech content related cues in a systematic manner. *Group Decision and Negotiation*, 13:61–79, 2004.
- [21] C. R. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.