Published in final edited form as:

Pattern Recognit. 2014 January 1; 47(1): . doi:10.1016/j.patcog.2013.06.013.

Salient and Non-Salient Fiducial Detection using a Probabilistic Graphical Model

C. Fabian Benitez-Quiroz*, Samuel Rivera*, Paulo F.U. Gotardo*, and Aleix M. Martinez*

Computational Biology and Cognitive Science Laboratory, The Ohio State University, 469 Dresse Laboratories, 2015 Neil Avenue, Columbus, Ohio, 43210

Abstract

Deformable shape detection is an important problem in computer vision and pattern recognition. However, standard detectors are typically limited to locating only a few salient landmarks such as landmarks near edges or areas of high contrast, often conveying insufficient shape information. This paper presents a novel statistical pattern recognition approach to locate a dense set of salient and non-salient landmarks in images of a deformable object. We explore the fact that several object classes exhibit a homogeneous structure such that each landmark position provides some information about the position of the other landmarks. In our model, the relationship between all pairs of landmarks is naturally encoded as a probabilistic graph. Dense landmark detections are then obtained with a new sampling algorithm that, given a set of candidate detections, selects the most likely positions as to maximize the probability of the graph. Our experimental results demonstrate accurate, dense landmark detections within and across different databases.

Keywords

Shape modeling; detailed face shape detection; face detection; probabilistic graphical model; landmark detection

1. Introduction

Shape detection is an important problem in computer vision and pattern recognition with applications in recognition, tracking, and classification, amongst others. The goal is to accurately detect the 2D position of specific shape landmarks, or *fiducials*, in an image. Some applications such as 3D reconstruction and the recognition of facial expressions require that the deformable shape be described by a *dense* set of *salient and non-salient* landmarks for satisfactory results [1, 2, 3]. Unfortunately, current detection algorithms are typically tailored to locating only a few salient landmarks [4, 5, 6]. Some exceptions are the 3-Dimensional Morphable Model (3DMM) [7] and the 3D model of [8] which find a dense set of face landmarks. However, these methods require a 3D database to construct the model and, thus, cannot be learned directly from an image collection.

© 2013 Elsevier Ltd. All rights reserved.

*Corresponding Author. benitezc@ece.osu.edu (C. Fabian Benitez-Quiroz), riveras@ece.osu.edu (Samuel Rivera), gotardop@ece.osu.edu (Paulo F.U. Gotardo), aleix@ece.osu.edu (Aleix M. Martinez).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

In this paper, we propose a novel statistical pattern recognition algorithm to accurately detect a dense set of salient and non-salient landmarks in an image. Our methodology does not require 3D object databases and can be used to design landmark detectors for different types of objects – *eg*, faces, hands, and structures in medical images. Our approach utilizes the fact that many object classes exhibit a homogeneous structure such that any detected landmark provides contextual information that facilitates the detection of the other landmarks. For instance, Fig.1(a) shows a classical example with human faces as objects of interest. While at first sight one may not perceive the 10 faces in this image, once a few fiducials have been detected (*eg*, an eye or the nose), the remaining facial parts become readily apparent. Thus, the location of a fiducial automatically provides information on where to find the others.

In our new method, the relationship between every pair of landmark positions is encoded by the edges of a probabilistic graphical model, where each node represents a landmark position and its local texture. The local texture information of salient landmarks allow them to be detected reliably, whereas non-salient landmark detection is unreliable from just the local texture. Fortunately, the reliable detections constrain the position of non-salient landmarks and vice versa. In addition, the coarsely localized non-salient fiducials aid estimation of other non-salient and misdetected landmarks. As a key result, our detection algorithm can robustly estimate the positions of fiducials in areas such as face cheeks, where a simple local feature detector would generally fail, Fig.1(b)–(c). Hence, the resulting algorithm can be used to estimate dense landmark maps in 2D images as in 3DMMs, without requiring prior 3D models.

Our proposed methodology is depicted in Figure 2. Using our graphical model (presented in Section 3), landmark detection amounts to maximizing the joint probability of the graph's nodes in an image. To accomplish this, we propose a sampling algorithm which selects the most likely fiducial positions given a set of candidate detections (Section 4), which resolves the classical computational complexity problem of pattern recognition algorithms that employ graphical models. This algorithm can deal with missing detections, false positives, and occlusions. In addition, we show how to augment the graph and use the original low-level detections to infer many more landmark coordinates in an incremental fashion. Our experimental results demonstrate accurate, dense landmark detections within and across different databases (Section 5).

2. Related Work

Algorithms such as Active Appearance Models (AAM) [9] and 3DMM [7] use a probabilistic shape and texture model to interpret an object image. These models are learned from a set of annotated training samples so they cannot detect variations beyond what is specified in the training set. In addition, the global shape model often favors a configuration similar to the mean shape and fails to capture subtle important changes such as eye blinks or single eyebrow motion. Furthermore, these algorithms are best suited to subject-specific modeling. On the other hand, algorithms that rely on fiducial detections are able to fit salient key points reliably without being overconstrained by a global model. However, these approaches can yield unrealistic shape estimates when the global shape is not constrained. These methods have advanced considerably in recent years, with algorithms that even rival human manual annotations [10, 11, 12, 13, 14, 15]. However, they provide a very limited number of fiducials around salient features such as the eyes, nose, and mouth, and most require high-quality images.

Our new pattern recognition method overcomes the above shortcomings by utilizing the positive aspects of fiducial detection and probabilistic shape and texture model approaches.

We take advantage of advances on local feature detection to ensure that subtle shape changes are not missed by our method. Each landmark position takes into account the position of *all* other locally-detected fiducials to generate a plausible configuration for the whole set of detected points. If some landmarks cannot be detected or are misdetected by the local feature detector, the other fiducials will constrain estimation of their positions.

Graphical models have previously been used to guide the fiducial position estimation, although using different approaches which are limited to a very small number of landmarks. Felzenszwalb and Huttenlocher [4] use a tree structured graph to infer the location of 5 face fiducials. For tree based graphs, a poorly localized root node negatively influences all daughter nodes. In contrast, our model represents a dense interconnection of landmarks. Every node has influence from all other shape landmarks so the effect of a few poorly localized nodes is circumvented by information from other nodes in the graph. In another work, Everingham *et al.* [6] model the joint probability of 9 fiducial positions in faces using a mixture of Gaussian trees. Fiducial are found using a discriminative model with Haar-like features. This algorithm only detects stable points (salient features) and the graph is solely used for robustness to different poses. Gu and Kanade [5] use local feature detections to generate a set of candidate positions for each fiducial, and then select the most probable set of fiducials using a Bayesian model which encodes the object pose and shape. Their algorithm is initialized using an AAM and a set of fiducials are localized around each landmark. Our algorithm also relies on a set of local feature detections but the highly interconnected structure of our graph allows us to infer positions of non-salient features while being more robust to occlusions. In general the proposed method outperforms these probabilistic graphical models because it can infer potentially dozens of fiducial positions including non-salient ones. In addition, it easily extends to non-face objects.

3. Our Probabilistic Graphical Model

Many natural objects are highly structured. Human faces, for example, exhibit strong relationships between the positions of different parts of the face; one can estimate the right eye position reliably given the position of the left eye and the nose by symmetry. In this and other objects, the information from each detected fiducial can be used to infer all other landmarks including those which are unseen, misdetected, or poorly localized. Given this insight, an appropriate modeling scheme describes the pairwise relationship between all fiducials in an object as well as the global configuration. We model the affinity between two fiducials i and j using a potential function Φ_{ij} , and the global configuration using the potential function β . ensures that the global configuration is reasonable. The logical way to combine these sources of information is using a probabilistic graph.

We define the joint probability of p fiducial *positions* as,

$$P_{pos}(X) = \beta(X) \frac{1}{Z_{pos}} \prod_{i=1, j=1}^{i=p, j < i} \Phi_{ij}(x_i, x_j), \quad (1)$$

where X encodes the set of coordinates $x_1, \dots, x_p$, $x_i \in \mathbb{R}^2$ has the $2D$ coordinate of fiducial i , Z_{pos} is called the *partition function* which makes (1) behave as a probability density

function (pdf), and $\beta(X) = \exp\left(-\frac{\alpha}{2}(\hat{X} - \mu)^T \Sigma^{-1}(\hat{X} - \mu)\right)$ is the Mahalanobis distance from the translated and scale invariant shape X to the mean shape, μ . To obtain X , we

centered the shape to the origin and then normalize by its Frobenius norm. $\mu = \frac{1}{N} \sum \hat{X}_i$ and $\Sigma = \frac{1}{N} \sum (\hat{X}_i - \mu)(\hat{X}_i - \mu)^T$ are the mean and the covariance matrix of the training

samples. The parameter $\alpha \in [0, 1]$ controls the penalty of differing from the mean shape. Intuitively, shapes with less deformations prefer a higher value of α , while shapes with larger deformations favor a smaller α .

Assuming that the object is detected and scaled to a standard size, the displacement between two fiducials can be modeled as a bivariate normal distribution. Although displacement between different pairs of fiducials may vary in scale, the correlations may be the same. Therefore, it is important to use a normalized distance, such as the Mahalanobis distance, when measuring the displacement. Thus, the *potential functions* $\Phi_{ij}(\cdot, \cdot)$ are defined as

$$\Phi_{ij}(x_i, x_j) = \exp \left(-(1 - \alpha) \bar{w}_{ij} (\Delta_{ij} - \mu_{ij})^T \Sigma_{ij}^{-1} (\Delta_{ij} - \mu_{ij}) \right), \quad (2)$$

where w_{ij} is the weight of the edge between fiducials x_i and x_j , $\Delta_{ij} = (x_i - x_j) \in \mathbb{R}^2$ is the 2D pairwise distances between landmarks i and j for a particular shape, and the parameters μ_{ij} and Σ_{ij} are the sample mean and sample covariance of Δ_{ij} which are estimated from the training data. The relationship between landmark positions is encoded as pairwise distances to make the model translation invariant. Since the quadratic term in $\Phi_{ij}(\cdot, \cdot)$ is a Mahalanobis distance, it is maximum when $\Delta_{ij} = \mu_{ij}$ and monotonically decreases from there.

In the case of face shapes, the edge connecting the eye to the eyebrow should have a larger weight than the edge connecting the eye to the mouth because the eye more strongly constraints the position of the eyebrow than the mouth. Therefore, we scale the Mahalanobis distances by normalized positive scalar edge weights $w_{ij} \in [0, 1]$, to account for this type of

variation. The normalized edge weights are defined as $\bar{w}_{ij} = \frac{w_{ij}}{\sum_{k=1}^p \sum_{l=1}^{l < k} w_{kl}}$, where the edge weights w_{ij} specify the relative importance of the edges in the graph.

Because the graph encodes the pairwise structure of the shape, an edge connecting nodes i and j should have a large weight, w_{ij} , if knowing the position of node i strongly constrains the position of node j . Also, a large w_{ij} means the relationship between fiducials i and j will be emphasized relative to pairs with smaller weights in (2). That is because a larger w_{ij} forces the optimization algorithm to put more effort into bringing the pairwise distance Δ_{ij} closer to its mean value, μ_{ij} , to maximize the potential. The degree to which i constrains j is specified by the sum of the eigenvalues of the covariance matrix Σ_{ij} because the eigenvalues describe the variance along the principal axes of the joint distribution $\mathcal{N}(\mu_{ij}, \Sigma_{ij})$. A larger variance means we know less about the relative positions of fiducials; the distance between connected fiducials can take on a very wide range of values. However, a smaller variance means the relative position of fiducial j can be inferred with more certainty. Therefore, we

set $w_{ij} = \frac{1}{\|\Sigma_{ij}\|_F}$ by noting that $\|\Sigma_{ij}\|_F$ equals the square root of the sum of the eigenvalues of Σ_{ij} .

The graph described by equation (1) is sufficient to constrain shape estimates to be reasonable, but it is also important to consider the local texture. This is achieved by weighing the pdf by the probability that each landmark is correctly localized, where each texture implies a probability of correct localization. This texture weighting implies that the textures for each fiducial are assumed to be conditionally independent. The graph describing the joint pdf of the landmark *positions and textures* is defined as,

$$P(X) = \frac{1}{Z} \beta(X) \prod_{i=1, j=1}^{i=p, j < i} \Phi_{ij}(x_i, x_j) \prod_{k=1}^p \gamma_k(x_k), \quad (3)$$

where $\phi(\cdot)$ is a potential function of the i^{th} fiducial's local texture. More formally, these functions ϕ_i are the normalized confidences of the local detection. We calculate that confidence using a kernel based density estimation algorithm [16], and normalize such that $\phi(\cdot)$ behave as a probability.

4. Testing Procedure

4.1. Fiducial Detection

Fiducial position estimation relies on a set of local detections. The local detections are a result of evaluating a classifier for each fiducial at each image patch within the image regions expected to contain the associated fiducial. The regions are determined by the training data. The classifiers rely on features extracted from the image patches, and are based on the local texture or *context* features. Salient points like corners of the eyes prefer local texture (pixel value features) as in [10], while non-salient points such as points in the cheeks require a more global texture feature [17]. The classifiers are learned using Kernel Linear Discriminant Analysis (KLDA) [18, 19] and the annotated training data. KLDA is a pattern recognition technique that finds a low-dimensional projection of the image features $\mathbf{x} \in \mathbb{R}^d$ which maximally separates samples from different classes while grouping members of the same class. We use the Radial Basis Function (RBF) kernel defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right), \text{ where } \sigma^2 \text{ is a parameter to select, and } \mathbf{x}_j \text{ are the vectorized sample images. The kernel parameter is optimized using the criterion of [20].}$$

Each local detector returns a set of candidate positions $D_i = \{\hat{\mathbf{x}}_i^1, \dots, \hat{\mathbf{x}}_i^{m_i}\}$, where $\hat{\mathbf{x}}_i^j$ is the j^{th} candidate position of fiducial i , and m_i is the number of candidate detections for fiducial i . These candidate detections consisting of their position and corresponding texture provide information about the unknown true fiducial positions in the graph. The main task of this model is to find the set of candidate detections that maximize the graph probability, given their position and local texture. More formally, the objective is

$$X^* = \arg \max_{x_1, \dots, x_p} P(X), \text{ s.t. } x_i \in D_i, i=1 \dots p. \quad (4)$$

4.2. Estimation of the Fiducials

Estimating fiducial positions amounts to optimizing the objective function of equation (4). The size of our search space depends on the number of candidates per fiducial. This space is

very large and evaluating the probability of each configuration requires $\prod_{i=1}^p m_i$ trials. In the ideal case, we would like to evaluate the probability of all possible combinations. This is computationally intractable, so we generate a reduced set of probable configurations from the full set of candidates. Instead of uniformly testing candidate configurations, we generate configurations that are highly probable given our probabilistic model. This allows us to find a reasonable shape without testing every possible case. The most likely configuration initializes the second stage, a belief propagation type algorithm that fine tunes the estimate to maximize the proposed model.

To find a set of likely fiducial configurations, the optimization algorithm assumes that at each iteration, $p-1$ "known" fiducials (salient and non salient) are coarsely detected. We then randomly sample for the p^{th} fiducial knowing that its pdf is conditioned by the "known" ones. This process is iterated for all fiducials until a likely global configuration is found. More formally, to find the $\hat{\mathbf{x}}_i^{j_i}, i=1 \dots p$, which maximizes (3), we initialize x_i

randomly to one of the local detections in D_i , $i = 1 \dots p$. In the *first stage*, we sequentially update x_k , $k = 1 \dots p$, by taking a random sample from the conditional pdf of x_k given all other nodes:

$$P(x_k | x_{-k}) \propto \gamma_k(x_k) \exp \left(-\frac{1-\alpha}{n} \sum_{j \neq k} \bar{w}_{kj} (\Delta_{kj} - \mu_{kj})^T \Sigma_{kj}^{-1} (\Delta_{kj} - \mu_{kj}) + \frac{\alpha}{2} x_k^T \Sigma^{-kk} \hat{x}_k \right), \quad (5)$$

where $x_{-k} = \{x_1, x_2, \dots, x_{k-1}, x_{k+1} \dots x_p\}$, x is the k^{th} fiducial of X , $\Delta_{kj} = x_k - x_j$ and Σ^{-ij} is the inverse of the submatrix that measures the covariance of the i^{th} and j^{th} fiducial of scale and translation invariant shape X . This procedure is repeated for all the fiducials x_k to generate a sample X_i , where $i = 1, \dots, \text{maxIter}$. Note that the conditional pdf of equation (5) can be calculated explicitly and normalized for the set of discrete candidate positions. We store the final configuration and the corresponding probability from equation (3) after every iteration.

To be robust to empty and false positive candidates resulting from occluded or noisy images, we use the probabilistic graph of equation (1) to augment the sets D_i with the maximum-likelihood (ML) estimate of each fiducial position x_i given x_{-i} . Taking the logarithm of equation (1) yields a quadratic function. Taking derivatives, we find that

$$x_k^{MLE} = ((1-\alpha)G_p + \alpha G_s)^{-1} ((1-\alpha)W_p + \alpha W_s) \quad (6)$$

where

$$G_p = \left(\sum_{j \neq k} \bar{w}_{kj} \Sigma_{kj}^{-1} \right)^{-1}, \quad G_s = s_c \Sigma^{-kk}, \quad s_c = \|X\|^2$$

$$W_p = \left(\sum_{j \neq k} \left[\bar{w}_{kj} \Sigma_{kj}^{-1} (x_j - \mu_{kj}) \right] \right), \quad W_s = s_c \left(\Sigma^{-kk} \mu_k + \sum_{j \neq k} \Sigma^{-kj} (\hat{x}_k - \mu_k) \right).$$

μ_k is the k^{th} coordinate of $\mu + t_x$, where t_x is the centroid of X . Note that this procedure without taking into account x_k^{MLE} is commonly known as a Gibbs Sampling [21].

In the *second phase* of the optimization, we initialize with the previously sampled configuration X_i that maximizes equation (3), and repeat the sequential procedure until convergence or a maximum number of iterations. However, instead of drawing random samples from the conditional pdf, we select the value $x_k \in D_k$ which maximizes equation (3). The procedure is summarized in algorithm 1.

Algorithm 1**Inference Algorithm**

```

Input= $\{D_1, \dots, D_p\}$ 
for  $i = 1$  to  $p$  do
    Set  $x_i^0$  = random sample of  $D_i$ 
end for
for  $stage = sampling$  to  $belief propagation$  do
    for  $t = 1$  to  $maxIter$  do
        for  $i = 1$  to  $p$  do
            Set  $\tilde{D}_i = \{D_i, x_i^{MLE}\}$ 

            Calculate  $P(x_i^{t-1} | x_{-i}^{t-1})$  using (5) and the candidates  $D_i$ 
            if  $stage = sampling$  then
                Let  $x_i^t$  be a random sample of  $P(x_i | x_{-i})$ 
            else
                Let  $x_i^t$  be  $\arg \max_{x_i^j \in \tilde{D}_i} P(x_i^{t-1} | x_{-i}^{t-1})$ 
            end if

            Let  $x_i^{t-1} \rightarrow x_i^t$ 
        end for
        if  $x'$  did not change then
            end belief propagation
        end if
    end for
    Set  $x^0$  as the  $x^i$  that maximize (3)
end for

```

---

5. Experiments

We performed a series of experiments on face, heart, and hand shape detection to demonstrate the utility of the proposed approach. The parameter λ which controls the amount of regularization was cross validated separately for the different object types. The parameter $maxIter$ determines how many exemplars are obtained in the first stage of inference. We found experimentally that $maxIter = 100$ was sufficient to initialize the belief propagation near a local minimum of (3).

5.1. Face landmark detection

We first evaluate the proposed algorithm on three face databases: AR [22], Labeled Faces in the Wild (LFW) [23], and the XM2VTS database [24]. Faces are roughly localized in position and scale using the Viola-Jones face detector, then cropped and scaled to 150×150 pixels, corresponding to a mean inter-eye distance of approximately 15 pixels. For the AR

database, we train with 448 images, and test on 448 images containing subjects not found in the training set. For the LFW database, we train with 1027 images, and test on 500 images of subjects not used in training. For the XM2VTS database, we train with 448 images, and test on 350 images containing subjects not found in the training set. We also use the model trained with the LFW database to detect fiducials on the AR and XM2VTS databases. The error is measured as the Euclidean distance from the ground truth fiducial positions (*i.e.*, manual markings) to the estimated position for a total of 50 fiducials over all test images. Results are compared to those of the AAM algorithm.

Example detections using our algorithm are shown in Fig. 3 with error histograms shown in Fig. 4. The first three histograms are for within database testing. That is, the training and testing set, although disjoint, are from the same database. The last three histograms are for across database experiments. In this case, we employed the training set of one database and do the testing on the images of the other specified database. These are the most challenging and interesting experiments. Figure 5 shows an additional visual comparison between our method, [9] and [6].

Our algorithm outperforms the AAM in every case. For the case of constrained databases such as the XM2VTS and AR databases, our algorithm slightly outperforms the AAM. However, for the case of the unconstrained LFW database, and the across-database experiments, our algorithm performs significantly better than the AAM. This reiterates the problem of learning a global probabilistic shape and texture model. It also verifies the benefit of using a probabilistic graphical combined with local detection of fiducials.

We also compare our method to the algorithm of [6] on the AR and LFW databases for the 9 fiducial points detected by their algorithm. The implementation was obtained from the authors' website, and was already trained. On the AR database, our algorithm achieves an error with mean and standard deviation of 1.5315 ± 1.2751 pixels while [6] achieves an error rate of 1.9189 ± 1.4619 pixels. On the LFW database, our algorithm achieves an error with mean and standard deviation of 2.5052 ± 1.6972 pixels, while [6] obtains an error rate of 2.2801 ± 4.1774 pixels. On the XM2VTS database, our algorithm achieves an error with mean and standard deviation of 1.5561 ± 1.7515 pixels, while [6] obtains an error rate of 1.7922 ± 1.8497 pixels. In summary our method provides more accurate detections and additional set of non-salient landmarks.

5.2. Cardiac MRI

To demonstrate the flexibility of our method, we now show detections of landmarks describing the epicardial and endocardial contours of the left-ventricle (LV) on cardiac magnetic resonance images (MRI). Our image database includes images of 8 subjects with a total of 160 images that were rescaled to 100×100 pixels. For each subject, 20 images depict the contraction and relaxation of the LV in short-axis view during a complete cardiac cycle. Each image was manually marked by a cardiologist with 22 landmarks around the LV. We randomly split the data into 100 images for training and 60 images for testing. Fig. 6(a) shows examples of detected landmarks and the approximate LV boundaries for 4 different subjects. Note that the detected contours correctly segment papillary muscles and trabeculae together with the interior blood pool of the LV. The boundaries were obtained by interpolating the Fourier transform of sequences of landmark coordinates represented as complex numbers, $l_i = x_i + y_i \sqrt{-1}$. Following the same comparison procedure of Section 5.1, our algorithm outperforms AAM with a mean error and standard deviation of 2.4049 ± 2.2949 pixels while AAM achieves an error rate of 5.0019 ± 3.5206 pixels. The error histograms are shown in Fig. 6(b).

5.3. Hand Shapes

To further challenge the algorithm, we detected 52 landmarks describing the hand contour. The image database of [25] contains 40 images of 4 subjects showing different hand shapes. We rescaled the images to 100×100 pixels, then for 3 folds of cross validation, we randomly selected 30 images for training and 10 images for testing. Fig. 7(a) shows examples of the detected landmarks and the approximate contour of the hand.

Following the same comparison procedure of Section 5.1 and combining the 30 testing results of all cross validation folds, our algorithm outperforms AAM with a mean error and standard deviation of 1.8805 ± 2.3707 pixels while AAM achieves an error rate of 3.5008 ± 3.0955 pixels. The error histograms are shown in Fig. 7(b).

5.4. Incremental Learning: Inferring Additional Landmarks

So far we have shown two cases where the algorithm can be used for detecting shapes having salient and non-salient landmarks. This is sufficient for most applications, but some require a shape representation with many more landmarks. Given the assumption that a set of localized fiducials constrains the position of other shape landmarks, we should be able to accurately predict the position of $m \gg p$ fiducials based on a localized subset of p fiducials obtained using the above approach. To achieve this, note that equation (6) only relies on the shape part of the graphical model (no reliance on texture). Therefore, we can learn the parameters for an m node graph from an annotated set of images, and simply estimate the position of each of the undetected fiducials as the MLE estimate of that fiducial position given the position of the p previously detected fiducials using equation (6). Fig. 8 shows the much denser shape detection achieved using the 50 fiducial detections of Fig. 3 and an augmented shape model. Note that no local feature detection or iterative sampling was done to infer the much denser set of landmarks.

6. Conclusion

We have presented a new statistical pattern recognition algorithm for detecting a dense set of salient and non-salient landmarks in images of a deformable object. This method exploits the fact that each landmark position constrains the position of other landmarks on an object. In our model, the pairwise relationships between landmarks is naturally encoded in the nodes and edges of a probabilistic graph. Given a set of candidate landmarks provided by local detectors, our algorithm selects the set of candidate detections that maximize the joint probability of the graph. Besides being more accurate, our method also provides dense detections of deformable shapes and is not restricted to faces. The face experiments show detection of significant local deformations of the eyes and mouth, while the hand experiments demonstrate precise detection with global deformation. Furthermore, results including training and testing across different datasets show that our method outperforms specialized, salient feature detectors as well as AAMs.

Acknowledgments

This research was partially supported by NIH grants R01-EY-020834 and R21-DC-011081.

References

1. Martinez AM, Du S. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *Journal of Machine Learning Research*. 2012; 31(1):1589–1608. [PubMed: 23950695]
2. Berger, M.; Ponroy, J.; Wrobel-Dautcourt, B. Realistic face animation for audiovisual speech applications: A densification approach driven by sparse stereo meshes; *Proceedings of Computer*

- Vision/Computer Graphics Collaboration Techniques 4th International Conference, MIRAGE 2009; 2009. p. 297-307.
3. Hamsici, OC.; Gotardo, PFU.; Martínez, AM. Learning spatially-smooth mappings in non-rigid structure from motion; Proceedings of the European Conference on Computer Vision; 2012. p. 260-273.
  4. Felzenszwalb PF, Huttenlocher DP. Pictorial structures for object recognition. International Journal of Computer Vision. 2005; 61(1):55–79.
  5. Gu, L.; Kanade, T. A generative shape regularization model for robust face alignment; Proceedings of the European Conference on Computer Vision; 2008. p. 413-426.
  6. Everingham, M.; Sivic, J.; Zisserman, A. Hello! my name is... buffy - automatic naming of characters in tv video; Proceedings of the British Machine Vision Conference; 2006.
  7. Blanz, V.; Vetter, T. A morphable model for the synthesis of 3D faces; Proceedings of SIGGRAPH; 1999. p. 187-194.
  8. Gu, L.; Kanade, T. 3D alignment of face in a single image; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2006. p. 1305-1312.
  9. Cootes, TF.; Edwards, GJ.; Taylor, CJ. Active appearance models; Proceedings of the European Conference on Computer Vision; 1998. p. 484-498.
  10. Ding L, Martinez AM. Features versus context: An approach for precise and detailed detection and delineation of faces and facial features. IEEE Trans. on Pattern Analysis and Machine Intelligence. 2010; 32:2022–2038.
  11. Moriyama T, Kanade T, Xiao J, Cohn JF. Meticulously detailed eye region model and its application to analysis of facial images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006; 28(5):738–752. [PubMed: 16640260]
  12. Zhou, F.; De la Torre, F.; Cohn, J. Unsupervised discovery of facial events; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2010.
  13. Li, P.; Prince, SJD. Joint and implicit registration for face recognition; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Miami, FL, USA. 2009. p. 1510-1517.
  14. Sznitman R, Jedyak B. Active testing for face detection and localization. IEEE Trans. on Pattern Analysis and Machine Intelligence. 2010; 32:1914–1920.
  15. Rivera S, Martínez AM. Learning deformable shape manifolds. Pattern Recognition. 2012; 45(4): 1792–1801. [PubMed: 22308002]
  16. Gray, A.; Moore, E. Very fast multivariate kernel density estimation using via computational geometry; Proceedings of the Joint Stat. Meeting; 2003.
  17. Shechtman, E.; Irani, M. Matching local self-similarities across images and videos; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2007.
  18. Mika, S.; Ratsch, G.; Weston, J.; Scholkopf, B.; Muller, K. Fisher discriminant analysis with kernels; Proceedings of the IEEE Neural Networks for Signal Processing Workshop; 1999. p. 41-48.
  19. Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach. Neural Computation. 2000; 12(10):2835–2404.
  20. You D, Hamsici O, Martinez AM. Kernel optimization in discriminant analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence. 2011; 33(3):631–638.
  21. Geman, S.; Geman, D. Readings in computer vision: issues, problems, principles, and paradigms. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1987.
  22. Martínez A, Benavente R. The AR face database. Tech. Rep. 24, Computer Vision Center. 1998 Jun.
  23. Huang, B.; Ramesh, M.; Berg, T.; Learned-Miller, E. Tech. Rep. 07–49. University of Massachusetts - Amherst; 2007 Oct. Labeled faces in the wild: A database for studying face recognition in unconstrained environments.
  24. Messer, K.; Matas, J.; Kittler, J.; Jonsson, K. Xm2vtsdb: The extended m2vts database; Proceedings of the Second International Conference on Audio and Video-based Biometric Person Authentication; 1999. p. 72-77.

25. Stegmann, MB.; Gomez, DD. Tech. rep. Technical University of Denmark; 2002. A brief introduction to statistical shape analysis.

Biographies

C. Fabian Benitez-Quiroz received his BS degree in Electrical Engineering from the Pontificia Universidad Javeriana, Cali, Colombia, in 2004 and his MS in Electrical Engineering from University of Puerto Rico in 2008. He is currently working toward his PhD degree in Electrical Engineering as a member of the Computational Biology and Cognitive Science Lab at OSU. His research interests include analysis of facial expressions in sign languages, functional data analysis, deformable shape detection, and face perception.

Samuel Rivera received his BE degree in Electrical Engineering from the University of Delaware in 2007, his MS in Electrical Engineering from The Ohio State University (OSU) in 2012, and his PhD degree in Electrical Engineering as a member of the Computational Biology and Cognitive Science Lab at OSU in 2012. His research interests include deformable shape detection, machine learning, category learning, and emotion perception in infants and adults.

Paulo F.U. Gotardo is a Postdoctoral Researcher with the Computational Biology and Cognitive Science Lab at the Department of Electrical and Computer Engineering, The Ohio State University. His research interests include machine vision and learning, with a focus on the 3D reconstruction of dynamic scenes and human-computer interaction. His prior research experience also includes range image understanding and the analysis of ventricular contraction patterns in cardiac MRI. He received the BSc and MSc degrees in Informatics from Universidade Federal do Parana, Brazil, in 2000 and 2002, and the PhD degree in Electrical and Computer Engineering from The Ohio State University in 2010. During his graduate studies, he was a recipient of excellence fellowships from The Brazilian Ministry of Education and The Fulbright Committee.

Aleix M. Martinez is an associate professor in the Department of Electrical and Computer Engineering at The Ohio State University (OSU), where he is the founder and director of the Computational Biology and Cognitive Science Lab. He is also affiliated with the Department of Biomedical Engineering and with the Center for Cognitive Science. Prior to joining OSU, he was affiliated with the Electrical and Computer Engineering Department at Purdue University and with the Sony Computer Science Lab. He is an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence and of Image and Vision Computing, and has served as area chair at CVPR, ICCV, ICPR and F&G. His areas of interest are learning, vision, linguistics, and their interaction.

Highlights

- The new pattern recognition method accurately detects a dense set of object landmarks in an image.
- The relative positions of pairs of landmarks are encoded by a probabilistic graph.
- Non-salient landmarks, missed by other detectors, are now robustly detected.
- It is applicable to different objects e.g., faces, hands, and cardiac structures.

(a) From salient to non-salient features (b) Standard detectors (c) Our dense detection

Figure 1.

Can you find the 10 faces in (a)? These faces are difficult to see until a face feature is detected (*eg*, a nose); then the entire face becomes salient. (b) The output of a standard face landmark detector is typically restricted to a few salient points. (c) Our novel method provides dense detections that include both salient and non-salient landmarks.

Figure 2.

Illustration of proposed methodology: The 'Graph Model' block shows the graph learning phase of the method (Section 3), where we model the relationship between salient and non-salient landmarks as a probabilistic graph. Thicker edges represent larger weights between fiducials. Only a subset of edges of the fully connected graph are shown. The 'Test Image' block highlights cases of misdetections and multiple detections for a particular fiducial. The 'Shape Detection' block (section 4) shows that the graph model is used along with the local feature detections to determine the most probable shape configuration.

Figure 3.

We show example results of the derived approach. From top to bottom, the rows correspond to the shape detections for the AR database [22], the LFW database [23], and the XM2VTS database [24]. The database contains face images in unconstrained environments. Results show robustness to occlusions, pose, and lighting.

Figure 4.

Error histograms (Euclidean distance, in pixels) for a total of 50 detected fiducial points *versus* the ground truth of the testing sets. The ground truth positions were obtained by manual annotation. PGA denotes our new probabilistic graph algorithm, while AAM denotes the classical AAM.

Figure 5.

Comparison of our algorithm with AAM [9] and the local detector of [6]. Our method provides more precise detections than the AAM, and many more fiducial points than the local feature detector.

(a)

(b)

Figure 6.

Heart Experiments. (a) Detected landmarks delineating the epicardial and endocardial contours of the LV in cardiac MRI (best seen in color). (b) Error histogram (Euclidean distances, in pixels) for 22 detected landmarks *versus* the ground truth of the testing images.

(a)

(b)

Figure 7. Hand Experiments on [25]. (a) Detected landmarks delineating the shape of the Hand. (b) Error histogram (Euclidean distances, in pixels) for 52 detected landmarks *versus* the ground truth of the testing images.

Figure 8.

We show a denser set of fiducial positions which were inferred using the detections from Fig. 3 and an expanded probabilistic graph as in Section 5.4.

Algorithm 1**Inference Algorithm**

Input= $\{D_1, \dots, D_p\}$

for $i = 1$ to p **do**

    Set x_i^0 = random sample of D_i

end for

for $stage = sampling$ to $belief\ propagation$ **do**

for $t = 1$ to $maxiter$ **do**

for $i = 1$ to p **do**

            Set $\tilde{D}_i = \{D_i, x_i^{MLE}\}$

            Calculate $P(x_i^{t-1} | x_{-i}^{t-1})$ using (5) and the candidates D_i

if $stage = sampling$ **then**

                Let x_i^t be a random sample of $P(x_i^t | x_{-i}^t)$

else

                Let x_i^t be $\arg \max_{x_i^j \in \tilde{D}_i} P(x_i^{t-1} | x_{-i}^{t-1})$

end if

            Let $x_i^{t-1} \rightarrow x_i^t$

end for

if x^t did not change **then**

            end belief propagation

end if

end for

    Set x^0 as the x^i that maximize (3)

end for