Facial Point Detection with Occlusion Insensitive Visibility-Aware Part Model

Yuanqian Li, Yanfei Liu, and Xi Zhou

Chongqing Institute of Green and Intelligent Technology,
Chinese Academy of Sciences, 401122, Chongqing, China
{liyuanqian,liuyanfei,zhouxi}@cigit.ac.cn

Abstract. In this paper, we describe a method named visibility-aware part model for facial point detection in static images based on the pictorial structure model. A binary part visibility term is introduced to describe the occlusion state of each part, which can determine which facial points are occluded. The introduction of the term enhances the representation power of the model especially for the occlusions. The combining of the structure constrains and the powerful appearance model makes the model more robust and reduces the possibility of model crashing in some extent. Experimental results show that our proposed model can detect facial feature points accurately and robustly under occlusions.

1 Introduction

The facial feature points are the prominent landmarks around facial component such as eyebrows, eyes, nose, mouth and face contour. Facial feature point detection plays foundational roles in many face analysis tasks, such as face recognition, pose estimation and 3D face reconstruction, etc. In these tasks, Accurate and efficient facial feature point detection is desirable to make automated face analysis systems more effective.

Existing methods for facial feature point detection can be categorized into three categories: texture-based, shape-based and methods combining texture and shape.

Texture-based methods model the local texture around a given feature point, for example the pixel values in a small region around an eye corner. Typical texture-based methods include Eigenfaces [1], Elastic Bunch Graph Matching (EBGM) [2], and Local Feature Analysis (LFA) [3], etc. Shape-based methods regard all facial feature points as an holistic shape, which is learned from a set of labeled faces, and try to find the proper shape for any known face. Typical shape-based methods include Active Shape Models (ASM) [4], Active Appearance Models (AAM) [5] and 3D Morpharable Model [6]. The researchers also proposed methods combining texture and shape information. For example, [7] uses principal component analysis (PCA) on the grey level images combined with ASM, [8] extends the AAM with Constrained Local Models, and [9] applies a boosting algorithm to determine facial feature point candidates for each pixel in an input image and then uses a shape model as a filter to select the most probable position of five points.

B. Huet et al. (Eds.): PCM 2013, LNCS 8294, pp. 25–34, 2013.

Fig. 1. Illustration of typical occlusion conditions (images from the LFW database [10])

Overall, most of the existing methods are only feasible under conditions without occlusion, and few previous works have reported to be able to robustly handle large occlusions such as glasses, beards, and hair which partially cover the eyes or the mouth. As illustrated in Figure 1, occlusion is a common phenomenon in the real world. However, it is very challenging for facial feature point detection, and makes many face processing tasks based on feature point detection cumbersome, such as face recognition. On one hand, even the most sophisticated feature point detection model will crash under the condition of partial occlusions. Consequently the points without occlusion cannot be precisely detected, which seriously affects the robustness of the detection model. On the other hand, assuming the points without occlusion are detected accurately, location prediction of the occluded points is also a tough problem, although it is usually useful to many face processing tasks. Given these issues, accurate and efficient detection of facial feature points under partial occlusions remains challenging.

Pictorial structure model is an influential approach in object recognition, which decomposes the appearance of an object into local part templates, together with geometric constraints on pairs of parts [11]. Inspired by the pictorial structure model, we introduce a novel representation for modelling part-occluded faces, Visibility-aware Part Model (VPM), to address the occlusion problem. Unlike traditional models for object recognition using parts parameterized solely by location, we introduce a visibility state for each part to describe the occlusion state of the part, which can determine which of the facial points are occluded. Firstly, the introduction of the visibility state makes the appearance model more powerful to represent the occlusions, which allows our proposed model to predict the occluded points more reliable. Secondly, since our model is motivated by the pictorial structure, in our model the feature point is defined to be a spatial relation between two facial feature points rather than being a facial feature point itself. The model combining of the structural constrain and the appearance model of powerful representation ability will automatically produce corrected point configurations that preserve well-estimated points. Extensive experiments on facial point detection show that the proposed model can localize facial feature points accurately even under severe occlusions.

The paper is structured as follows. Section 2 proposes our visibility-aware part model. Section 3 presents the algorithm for learning model parameters and the method for fitting the learned model to test images. Section 4 discusses the experimental results. The paper concludes in Section 5.

2 Model

In the Pictorial Structure an object is first decomposed into parts and then the best part candidates are searched subject to some spatial constraints such that the likelihood of generating the concerned image is maximized. The pictorial structure model can usually be expressed as an N-node relational but undirected graph $G=(Pt,E)$, where $Pt=\{Pt_1,Pt_2,...,Pt_N\}$ corresponds to the N parts and the edges $E=\{(Pt_i,Pt_j),i\neq j\}$ specify which pairs of parts have consistent relations. Let θ to be a set of parameters that define an object model, I denote an image, and $L=(l_1,l_2,...,l_N)$ denote a configuration of the object, where each l_i specifies the location of each part Pt_i on the image plane. To infer the locations of the parts of an object from this model, we can search for the maximum a posterior $p(L|I,\theta)$, i.e., the probability that a face configuration is L given the model θ and an image I. Using Bayes rule, the posterior can be written as

$$P(L|I,\theta)\propto p(I|L,\theta)p(L|\theta) \tag{1}$$

where $p(I|L,\theta)$ is the generative model of the appearance and $p(L|\theta)$ measures the prior probability that a face appears at the location L. In general, the model parameter is denoted by $\theta=(u,c)$, where u expresses the appearance while c expresses structural constraints on edges. This provides the opportunity to predict the location of occluded points by reconstructing new model. Many objects, including faces and people can be represented by such multi-part models in its simplicity and generality, thus, the pictorial structure formulation is appealing in many object recognition fields [12-15].

Specially in our facial feature point detection task, we model every facial feature point as a part Pt_i. For the purpose of precise detection of feature points under occlusions, we introduce a binary visibility state term $s=\{s_1,s_2,\cdots,s_N,\}$, $s_i\in\{0,1\}$ to present whether the part is occluded or not, where $s_i=1$ denotes the part is visible while $s_i=0$ denotes the part is not visible or occluded. Thus, the posterior can be written as $P(L,s|I,\theta)$ in our model,

$$P(L,s|I,\theta)\propto p(I|L,s,\theta)p(L,s|\theta) \tag{2}$$

Our visibility-aware part model is parameterized by $\theta=(u,c)$, where $u\in\{u_1,u_2,\cdots,u_N\}$ are appearance parameters, and $c=\{c_{ij}\mid(Pt_i,Pt_j)\in E\}$ are connection parameters which denote the spatial relationships between parts. Obviously, the first term in Eq. (2) depends only on the appearance of the parts, while the second term depends only on the connection parameters. Assuming that the parts are statistically independent for appearance, we have

$$p(I\mid L,s,\theta)=p(I\mid L,s,u)=\prod_{i=1}^{N}p(I\mid l_i,s_i,u_i) \tag{3}$$

The appearance of each part can be modelled by unimodal Gaussian distribution $p(I\mid l_i,u_i)\propto\mathcal{N}\left(\alpha(l_i),\mu_i,\Sigma_i\right)$, where $u_i=(\mu_i,\Sigma_i)$ and $\alpha(l_i)$ is a high-dimensional feature vector of an image patch centered at the position l_i in [13], which collects all the responses of a set of filters of different scales at the point l_i. Then we have

$$\prod_{i=1}^{N}p(I\mid l_i,s_i,u_i)=\prod_{i=1}^{N}(s_i\cdot\mathcal{N}\left(\alpha(l_i),\mu_i,\Sigma_i\right)+(1-s_i)\cdot\mathcal{N}(\alpha(l_i),\mu_i^{'},\Sigma_i^{'})) \tag{4}$$

And in our model, $u_i=(\mu_i,\Sigma_i,\mu_i^{'},\Sigma_i^{'})$, (μ_i,Σ_i) is the appearance parameter of unoccluded point and $(\mu_i^{'},\Sigma_i^{'})$ denotes the appearance parameter of point under occluded state. The value of s_i determines which distribution is used as the detecting model. The appearance parameters $(\mu_i^{'},\Sigma_i^{'})$ can be learned from the information provided by occluded example images. Obviously, the introduction of s_i enhances the representation ability of the appearance model and makes the model more robust to occlusions.

With a similar independent assumption on the edge constraints between components, we have the structure model

$$p(L,s\mid\theta)=p(L,s\mid c)=\prod_{(Pt_i,Pt_j)\in E}p(l_i,l_j,s_i,s_j\mid c_{ij}) \tag{5}$$

Similarly, the spatial relationships between pairs of parts can also be modelled by Gaussian distribution $p(l_i,l_j\mid c_{ij})\propto\mathcal{N}\left(l_i-l_j,\mu_{ij},\Sigma_{ij}\right)$ [13], where $c_{ij}=\left(\mu_{ij},\Sigma_{ij}\right)$. The spatial relationships between pairs of parts will not change no matter if the points are occluded or not. Therefore, the structure model has nothing relationship with s_i and s_j, and can then be rewritten as

$$\begin{aligned}&\prod_{(Pt_i,Pt_j)\in E} p(l_i,l_j,s_i,s_j \mid c_{ij})\\ &= \prod_{(Pt_i,Pt_j)\in E} p(l_i,l_j \mid c_{ij}) = \prod_{(Pt_i,Pt_j)\in E} \mathcal{N}\left(l_i - l_j, \mu_{ij}, \Sigma_{ij}\right)\end{aligned} \tag{6}$$

Plugging Eq. (3), Eq. (4), Eq. (5) and Eq. (6) into Eq. (2), we get the global objective function

$$\begin{aligned}&p(L,s \mid I,\theta) \propto \{\prod_{i=1}^{N} p(I \mid l_i,s_i,u_i) \prod_{(Pt_i,Pt_j)\in E} p(l_i,l_j,s_i,s_j \mid c_{ij})\}\\ &= \prod_{i=1}^{N} (s_i \cdot \mathcal{N}\left(\alpha\left(l_i\right),\mu_i,\Sigma_i\right) + \left(1-s_i\right)\cdot \mathcal{N}(\alpha\left(l_i\right),\mu_i^{'},\Sigma_i^{'}))\cdot\\ &\prod_{(Pt_i,Pt_j)\in E} \mathcal{N}\left(l_i - l_j, \mu_{ij}, \Sigma_{ij}\right)\end{aligned} \tag{7}$$

Facial feature points are frequently occluded by beard, hair, glasses and other objects such as tennis ball. Since occlusions often do not happen at random, the locations of occluded points (parts) may have consistent appearance. We model occlusions by learning separate appearance parameters $(\mu_i^{'},\Sigma_i^{'})$ for occluded points.

3 Learning and Inference

3.1 Learning Model Parameters

Suppose we have a set of M example images $\{I^1, I^2, \cdots, I^M\}$ including images with various face-partial occlusions and also images without any occlusion on the face. The corresponding feature point locations $\{L^1, L^2, \cdots, L^M\}$ and visibility states $\{s^1, s^2, \cdots, s^M\}$ are also labeled for these training images. By definition, the maximum likelihood estimate of θ is the value θ^* that maximizes $p(L^1,\cdots,L^M,s^1,\cdots,s^M \mid I^1,\cdots I^M,\theta)$. Assuming each example is generated independently, it can be rewritten as

$$\begin{aligned}\theta^* &= \arg\max_{\theta} \prod_{k=1}^{M} p(L^k,s^k \mid I^k,\theta)\\ &= \arg\max_{\theta} \prod_{k=1}^{M} p(I^k \mid L^k,s^k,\theta) \prod_{k=1}^{M} p(L^k,s^k \mid \theta)\end{aligned} \tag{8}$$

The first term in this equation depends only on the appearance of the parts, while the second term depends only on the connection parameters. As a consequence, we can use this framework as long as there is a maximum likelihood (ML) estimation procedure for learning the model parameters for a single part from training images.

For the appearance parameters u, we have

$$u^* = \arg\max_u \prod_{k=1}^{M} p(I^k \mid L^k, s^k, u) \tag{9}$$

From Eq. (3) and Eq. (4), we get

$$\begin{aligned} u_i^* &= \arg\max_{u_i} \prod_{k=1}^{M} p(I^k \mid l_i^k, s_i^k, u_i) \\ &= \arg\max_{\mu_i, \Sigma_i, \mu_i', \Sigma_i'} \prod_{k=1}^{M} (s_i^k \cdot \mathcal{N}\left(\alpha\left(l_i^k\right), \mu_i, \Sigma_i\right) + (1 - s_i^k) \cdot \mathcal{N}(\alpha\left(l_i^k\right), \mu_i', \Sigma_i')) \end{aligned} \tag{10}$$

This is exactly the ML estimation of the appearance parameters for part Pt_i, given independent examples $\{(I^1, l_i^1, s_i^1), (I^2, l_i^2, s_i^2), \cdots, (I^M, l_i^M, s_i^M)\}$. Similarly for the connection parameters c, we have

$$c^* = \arg\max_c \prod_{k=1}^{M} p(L^k, s^k \mid c) \tag{11}$$

From Eq. (5) and Eq. (6), we get

$$c_{ij}^* = \arg\max_{c_{ij}} p(l_i^k, l_j^k, s_i^k, s_j^k \mid c_{ij}) = \arg\max_{\mu_{ij}, \Sigma_{ij}} \prod_{k=1}^{M} \mathcal{N}\left(l_i^k - l_j^k, \mu_{ij}, \Sigma_{ij}\right) \tag{12}$$

This is the ML estimation for the joint distribution of l_i and l_j, given independent training examples $\{(I^1, l_i^1, s_i^1), (I^2, l_i^2, s_i^2), \cdots, (I^M, l_i^M, s_i^M)\}$. Learning our model involves picking labeled landmarks on a number of human face. Then using the ML estimation procedure, the appearance models for each part and spatial relationships between parts are automatically estimated from these training examples.

3.2 Inference

Inference corresponds to maximizing $p(L \mid I, \theta)$ from Eq. (6) given learned parameters $\theta = (u^*, c^*)$. Simply enumerate all locations and all values of s (0 or 1), and find the best configuration of parts.

$$
\begin{aligned}
L^* &= \arg\max_L p(L \mid I, \theta) \\
&= \arg\min_L \left(\sum_{i=1}^{N} -\log p(I \mid l_i, s_i, u_i) + \sum_{(Pt_i, Pt_j) \in E} -\log p(l_i, l_j, s_i, s_j \mid c_{ij})\right)
\end{aligned} \tag{13}
$$

Where $-\log p(I \mid l_i, s_i, u_i)$ is a match cost measuring how well part Pt_i matches the image data at location l_i, and $-\log p(l_i, l_j, s_i, s_j \mid c_{ij})$ is a deformation cost measuring how well the relative locations of the part Pt_i and Pt_j agree with the deformable model. In this work, we use a tree structure to construct the model. Therefore, the dynamic programming (DP) algorithm [13] can be adopted to find a location L with maximum posterior probability, which is used to detect facial feature points in novel images. We omit the message passing equations for lack of space.

4 Experiments

We firstly test how well the database copes with occlusions. Then, we perform a benchmark comparison of our proposed method with the existing state of the art.

The images are selected from LFW [10] and CIGIT (Chongqing Institute for Green and Intelligent Technology) databases to form a new database including images without occlusion and images with many occlusions, such as the mouth area occluded by beards and the eye areas occluded by glasses and hair, etc. Then this database is taken as the training set of our proposed algorithm. The CIGIT dataset is being collected by our team as a continual project, whose goal is to simulate the partially-controlled surveillance scenarios, including 91 poses, 6 facial movements, 4 kinds of occlusions, and 5 sets of combined indoor, controlled and outdoor, uncontrolled lighting. The LFW and CIGIT dataset not only have many occlusions, but also cover almost all the complex conditions in the real world. Thus, we select images from these two datasets to train our model.

To evaluate the precision of feature point detection, we adopt the most popular measure m_{e17} proposed by [8], which is a mean error over all internal points (17 points). And to compare our work with the current state of the art, we evaluated our method on the BioID database [16] which is one of the benchmark databases used by most facial point detection works. We compare our approach with the traditional pictorial structure model (PSM) [13], and the constrained local model (CLM) [8].

Figure 2 plots the cumulative error distribution curves of the compared methods, where the horizontal axis is the normalized Euclidean distance (m_{e17}), while the vertical axis is the cumulative localization score, showing the percentage of images that have been successfully processed corresponding to the detection error. As expected, our proposed method outperforms the traditional PSM and CLM.

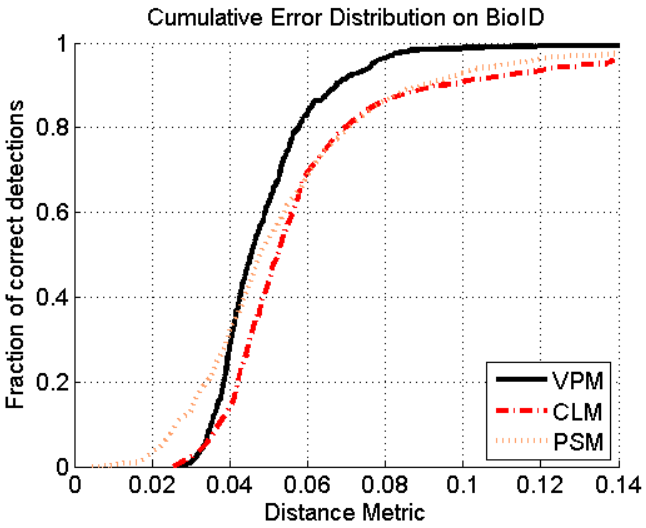

Fig. 2. Comparing the cumulative error distribution curves

Figure 3, Figure 4 and Figure 5 illustrate detection results on images with partially occluded faces. As shown in Figure 3, Figure 4, and Figure 5, the detection algorithm based on our model automatically handles partial occlusion in a robust way.

Fig. 3. Detection results on LFW database

Fig. 4. Detection results on BioID database

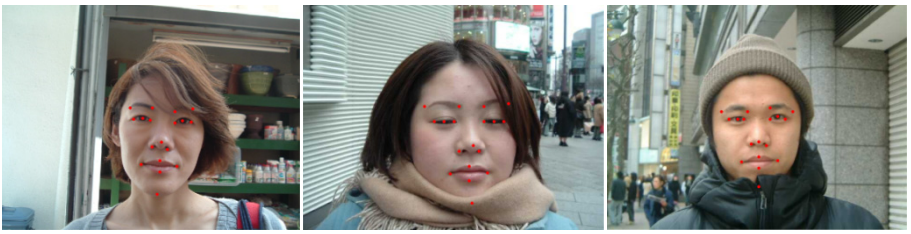

Fig. 5. Detection results on Yamaha database

To make this clear, we tabulate the percentage of successful localization subject to $m_{e17} < 0.05$ and $m_{e17} < 0.1$ in Table 1. It can be observed that the successful detection rate is promoted from at most 61.79% to 89.5% at $m_{e17} < 0.05$, and our algorithm achieves the best correct localization rate of 98.15% at $m_{e17} < 0.1$. Obviously, our method provides substantial performance compared to other methods.

Table 1. Percentages of successful detection to $m_{e17} < 0.05$ and $m_{e17} < 0.1$, respectively

Method	$m_{e17} < 0.05$	$m_{e17} < 0.1$
Traditional PSM	54.36%	92.91%
CLM	43.33%	90.89%
EASM	61.79%	97.32%
MKL-SVM	45.26%	92.5%
Our method	89.5%	98.15%

5 Conclusion

Aiming at resolving the problem of occlusion, we presented a model for facial point detection in this paper, which we refer to as the visibility-aware part model. Instead of modelling an object using parts parameterized solely by location, we introduce a "visibility" for each part to describe the occlusion state of the part. The introduction of the "visibility" provides us the accurate detection of facial feature points under occlusion. We show that the proposed visibility-aware part model can obtain better facial feature point detection results under partial occlusion. We also show that our proposed method outperforms the traditional PSM and CLM method when applied to the BioID data sets.

Acknowledgement. This research was supported by two projects from committee on science and technology of Chongqing, with Grant No. cstc2011ggC40009 and cstc2012gg-sfgc0079, respectively.

References

[1] Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
[2] Wiskott, L., Fellous, J.M., Kuiger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 775–779 (1997)
[3] Penev, P.S., Atick, J.J.: Local feature analysis: a general statistical theory for object representation. Network: Computation in Neural Systems 7, 477–500 (1996)
[4] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active Shape Models-Their Training and Application. Computer Vision and Image Understanding 61, 38–59 (1995)
[5] Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 681–685 (2001)
[6] Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 1063–1074 (2003)
[7] Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 504–513. Springer, Heidelberg (2008)
[8] Cristinacce, D., Cootes, T.: Feature detection and tracking with constrained local models. In: Proc. British Machine Vision Conference, pp. 929–938 (2006)
[9] Chen, L., Zhang, L., Zhang, H., Abdel-Mottaleb, M.: 3D shape constraint for facial feature localization using probabilistic-like output. In: Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 302–307 (2004)
[10] Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In: Workshop on Faces in 'Real-Life'Images: Detection, Alignment, and Recognition (2008)
[11] Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Transactions on Computers 100, 67–92 (1973)
[12] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1627–1645 (2010)
[13] Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International Journal of Computer Vision 61, 55–79 (2005)
[14] Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1385–1392 (2011)
[15] Felzenszwalb, P.F., Girshick, R.B., McAllester, D.: Cascade object detection with deformable part models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2241–2248 (2010)
[16] Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust face detection using the hausdorff distance. In: Audio-and Video-Based Biometric Person Authentication, pp. 90–95 (2001)