Active Pictorial Structures

Epameinondas Antonakos, Joan Alabort-i-Medina, Stefanos Zafeiriou
Department of Computing, Imperial College London.

Generic object alignment in terms of landmark points localization under unconstrained conditions (in-the-wild) is among the most challenging problems of computer vision. Significant research effort has been channelled towards developing deformable models with accurate performance and real-time computational cost. Two of the most well-studied deformable models are: (i) Pictorial Structures (PS) [4, 5, 9], and (ii) Active Appearance Models (AAMs) [3, 6]. In this paper, we propose Active Pictorial Structures (APS), a novel generative deformable model that takes advantage of the strengths and overcomes the weaknesses of both PS and AAMs. APS achieve state-of-the-art and close to real-time performance.

PS learn a patch expert for the appearance of each part of an object and model its shape using spring-like connections between landmarks based on a tree structure. Specifically, given an object class of n parts (landmarks) and a tree $G=(V,E)$, the cost function to be optimized is

$$\arg\min_{\mathbf{s}} \sum_{i=1}^{n} [\mathcal{A}(\ell_i)-\mu_i^a]^T(\Sigma_i^a)^{-1}[\mathcal{A}(\ell_i)-\mu_i^a]+ \\ + \sum_{i,j:(v_i,v_j)\in E} [\ell_i-\ell_j-\mu_{ij}^d]^T(\Sigma_{ij}^d)^{-1}[\ell_i-\ell_j-\mu_{ij}^d] \quad (1)$$

where $\mathbf{s}=[\ell_1^T,\ldots,\ell_n^T]^T$ is the vector of landmark coordinates ($\ell_i=[x_i,y_i]^T$, $\forall i=1,\ldots,n$) and $\mathcal{A}(\ell_i)$ is a feature vector extracted from the neighbourhood around the image location ℓ_i. $\{\mu_i^a,\Sigma_i^a\}$ and $\{\mu_{ij}^d,\Sigma_{ij}^d\}$ denote the mean and covariances of the appearance and deformation respectively. Inference is performed using a dynamic programming algorithm based on distance transform that can find a global minimum without any initialization. However, PS have two main disadvantages: (i) inference is very slow, and (ii) because the tree structure restricts too much the range of possible realizable shape configurations, the global optimum, even though it is the best solution in the span of the model, it does not always correspond to the shape that best describes the object in reality.

AAMs are generative models of the shape and appearance of an object. The shape model is built by applying Principal Component Analysis (PCA) on a set of aligned shapes. Similarly, the appearance model, which is represented in a holistic way (i.e. the whole texture is taken into account), is built by applying PCA on a set of shape-free appearance instances, acquired by warping the training images into a reference shape. Fitting AAMs involves solving a non-linear least squares problem and it is typically solved using a variant of the Gauss-Newton algorithm. AAMs have two disadvantages: (i) they are not fast enough for real-time applications, and (ii) by applying PCA the appearance of the object is modelled using a single multivariate normal distribution, which, as we show, restricts the fitting accuracy.

APS have a similar cost function as the one of PS (Eq. 1), which combines a shape and appearance model, similar to AAMs, along with a deformation prior. The biggest difference to both PS and AAMs is the use of Gaussian Markov Random Field (GMRF). By employing a GMRF, we make the assumption that the shape, appearance and deformation of the object can be modelled using multiple graph-based pairwise normal distributions between its parts. Specifically, given a graph $G=(V,E)$ and a set of abstract data vectors $\mathbf{x}_i$, $i=1,\ldots,|V|$, a GMRF formulates the precision matrix $\mathbf{Q}$ (i.e. inverse of covariance $\mathbf{Q}=\Sigma^{-1}$) of the data as a block-sparse matrix that has zeros at the blocks that correspond to disjoint vertexes, i.e. $\mathbf{Q}_{ij}=\mathbf{0}$, $\forall i,j:(v_i,v_j)\notin E$. Thus, APS consist of:

- A statistical linear shape model, similar to the one of AAMs, that is built based on an arbitrary undirected graph $G^s=(V^s,E^s)$. By applying PCA on the inverse precision matrix of the GMRF, we obtain an orthonormal basis $\mathbf{U}$ and a mean shape $\bar{\mathbf{s}}$. A new shape instance can be generated using the function $\mathcal{S}(\mathbf{s},\mathbf{p})=\mathbf{s}+\mathbf{U}\mathbf{p}$, where $\mathbf{p}$ are the shape parameters.
- An appearance model that is built based on an arbitrary undirected graph G^a and comprises of the mean appearance vector $\bar{\mathbf{a}}$ and the precision matrix $\mathbf{Q}^a$.
- A deformation prior, similar to the deformation part of the PS cost (Eq. 1), that is based on a directed graph G^d and includes the precision matrix $\mathbf{Q}^d$.

APS aim to minimize the cost function

$$\arg\min_{\mathbf{p}} \|\mathcal{A}(\mathcal{S}(\bar{\mathbf{s}},\mathbf{p}))-\bar{\mathbf{a}}\|^2_{\mathbf{Q}^a}+\|\mathcal{S}(\bar{\mathbf{s}},\mathbf{p})-\bar{\mathbf{s}}\|^2_{\mathbf{Q}^d}= \\ =\arg\min_{\mathbf{p}} [\mathcal{A}(\mathcal{S}(\bar{\mathbf{s}},\mathbf{p}))-\bar{\mathbf{a}}]^T\mathbf{Q}^a[\mathcal{A}(\mathcal{S}(\bar{\mathbf{s}},\mathbf{p}))-\bar{\mathbf{a}}]+ \\ +[\mathcal{S}(\bar{\mathbf{s}},\mathbf{p})-\bar{\mathbf{s}}]^T\mathbf{Q}^d[\mathcal{S}(\bar{\mathbf{s}},\mathbf{p})-\bar{\mathbf{s}}] \quad (2)$$

Inspired by AAMs, this problem is solved using a weighted inverse compositional Gauss-Newton optimization algorithm with fixed Jacobian and Hessian.

There are some important advantages of APS compared to PS and AAMs. (i) The proposed formulation allows to define any graph structure (not only tree) between the object's parts. This means that we can assume dependencies between any pair of landmarks for the shape, appearance and deformation, as opposed to PS that assumes independence for the appearance and a tree structure for the deformation. (ii) The sums of the cost function of PS in Eq. 1 are transformed into matrices multiplications in the cost function of APS in Eq. 2, which makes the computation much faster, especially in the case of objects with numerous parts. (iii) The spring-like deformation prior term of Eq. 2 makes APS much more robust than AAMs which lack such prior term. (iv) The inverse compositional Gauss-Newton optimization of APS has a close to real-time cost. Our Python implementation runs at $50ms$ per frame and is independent of the employed graph structures (G^s, G^a, G^d).

Our experiments on face alignment show that modelling the appearance of an object using a GMRF is much more beneficial than using PCA, as commonly done in the literature. Moreover, the deformation term makes the model robust in the case of bad initializations by restricting the shape model to generate only realistic shapes of the object. Finally, we show that APS, trained using a relatively small amount of training data, can compete and even surpass the accuracy of four of the most recently proposed state-of-the-art techniques in face alignment in-the-wild [2, 7, 8, 9] potentially trained with thousands of training examples. An open-source implementation of the proposed method is available as part of the Menpo Project [1] in `http://www.menpo.org/`.

[1] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proc. ACM MM*, pages 679–682, 2014.

[2] E. Antonakos, J. Alabort-i-Medina, G. Tzimiropoulos, and S. Zafeiriou. Hog active appearance models. In *Proc. ICIP*, pages 224–228, 2014.

[3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE T-PAMI*, 23(6):681–685, 2001.

[4] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE T-PAMI*, 32(9):1627–1645, 2010.

[6] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004.

[7] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *Proc. CVPR*, pages 1851–1858, 2014.

[8] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proc. CVPR*, pages 532–539, 2013.

[9] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. CVPR*, pages 2879–2886, 2012.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.