Nonparametric Facial Feature Localization

Birgi Tamersoy
birgi@utexas.edu

Changbo Hu
changbo.hu@gmail.com

J. K. Aggarwal
aggarwaljk@mail.utexas.edu

Computer and Vision Research Center
The University of Texas at Austin

Abstract

Any facial feature localization algorithm needs to incorporate two sources of information: 1) prior shape knowledge, and 2) image observations. Existing methods have primarily focused on different ways of representing and incorporating the image observations into the problem solution. Prior shape knowledge, on the other hand, has been mostly modeled using parametrized shape models. Parametrized shape models have relatively few parameters to control the shape variations, and hence their representation power is limited with the examples provided in the training data.

In this paper, we propose a novel method for modeling the prior shape knowledge. Rather than using a holistic approach, as in the case for parametrized shape models, we model the prior shape knowledge as a set of local compatibility potentials. This "distributed" approach provides a greater representation power as it allows for individual landmarks to move more freely. The prior shape knowledge is incorporated with local image observations in a probabilistic graphical model framework, where the inference is achieved through nonparametric belief propagation. Through qualitative and quantitative experiments, the proposed approach is shown to outperform the state-of-the-art methods in terms of localization accuracy.

Image credit: http://www.freshfaceclinic.com.au/

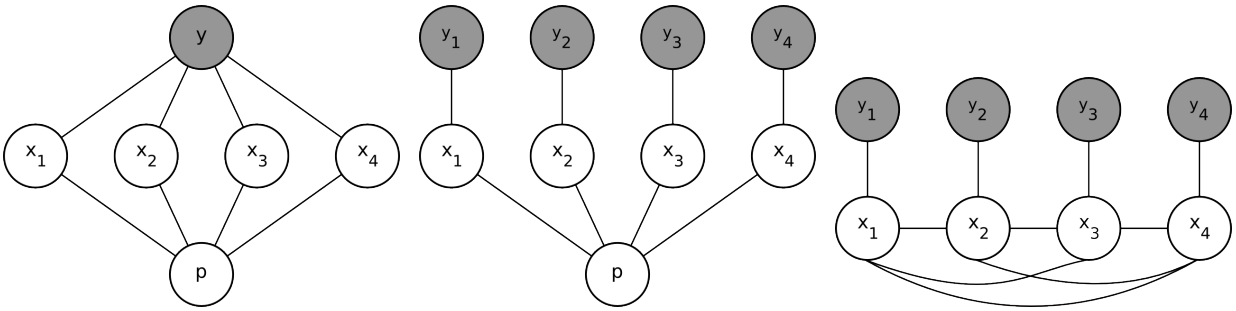
Figure 1. Example demonstrating the flexibility, generalization, and the accuracy of the proposed approach. The above result is obtained by fitting a *single* face model to the input image after *automatic* initialization. The local nature of the proposed method provides a great level of flexibility and fits both sides of the image equally well (best viewed in color and high-resolution).

1. Introduction

Facial feature localization is a crucial initial step in a wide variety of computer vision and human-computer interaction applications. Many algorithms require accurate feature locations as an input, and fail significantly once these features are slightly off. This heavy dependency on accuracy, combined with the evolving complexity of the interested applications, makes facial feature localization a popular and an important computer vision problem.

With more demanding applications, it becomes apparent that the current state-of-the-art solutions for facial feature localization are insufficient and need to be improved. Facial expression analysis is a good example. State-of-the-art feature localization algorithms perform adequately in the case of recognizing posed expressions, but they fell short in accuracy in the case of spontaneous expression recognition, where the feature dynamics are more subtle.

In this paper we propose a novel solution to the facial feature localization problem. Unlike most of the existing methods, we model the prior shape knowledge as a set of *compatibility potentials*, in a probabilistic graphical model. Local image observations are incorporated into this model as *observation potentials* at each node (i.e. landmark). The graph topology is determined automatically using training



(a) AAM-based. (b) CLM-based. (c) PGM-based (proposed).

Figure 2. Comparison of feature localization approaches in terms of how the prior shape knowledge and the image observations are modeled and incorporated. Empty nodes represent the landmark locations (and shape parameters), filled nodes represent the observations. The figure highlights the holistic nature of AAM- and CLM-based approaches vs. the local nature of the proposed PGM-based approach (note that AAM- and CLM-based approaches are illustrated as "graphical models" even though they are not PGM-based approaches).

data. Facial feature localization, or inference on the graph, is then achieved through nonparametric belief propagation. The result is a very flexible and easily expandable probabilistic framework for facial feature localization.

The paper is organized as follows: in Section 2 the related work is reviewed. Section 3 first provides the required technical background, and then introduces the proposed approach. Experimental results, comparing the proposed method to the state-of-the-art methods, are presented in Section 4. The paper is concluded with a discussion and mention of future work in Section 5.

2. Related Work

Majority of facial feature localization (or "model fitting") algorithms employ a parametrized shape model [5, 4, 7, 20, 27, 23]. Cootes and Taylor [5] coined the term "Point Distribution Model (PDM)" for this. PDM models the non-rigid shape variations of an object linearly and composes it with a global similarity transform:

$$x_i = sR(\bar{x}_i + \Phi_i q) + t \tag{1}$$

where x_i denotes the 2D-location of the PDM's i^{th} landmark, and $p = \{s, R, t, q\}$ are the PDM parameters consisting of a global scaling s, a rotation R, a translation t, and a set of non-rigid shape parameters q. Here, $\bar{x}_i$ is the mean location of the i^{th} landmark, and Φ is the matrix of basis variations (i.e. "shape vectors"). Shape vectors are obtained by performing a principle component analysis (PCA) on the training data.

PDM-based methods may further be categorized into two classes depending on how the image observation is incorporated into the shape model. The first class of methods, derived from Active Appearance Models (AAMs) [4], use a holistic "error image" to determine the parameter updates [4, 20]. The second class of methods, derived from Active Shape Models (ASMs) [5], are collectively named Constrained Local Models (CLMs) and utilize an independent set of local detectors. CLMs primarily differ on how the corresponding noisy response maps are approximated. Saragih et al. [23] show that a nonparametric Gaussian kernel density estimate (KDE) [24] of the response maps outperforms the existing parametric estimates [5, 27, 14].

PDM is a simple and an efficient method for modeling the deformations of objects, such as a human face. However, it is a fairly strict model where the representation power is limited with the shape variations presented in the training data.

Contrary to these strict parametrized models, relatively looser "part-based" statistical shape models have been proposed in the literature as well (e.g. [12, 30, 11]). In these statistical models the localization problem is formulated as finding the best configuration of the parts of the model ($\mathcal{L} = \{l_1, \ldots, l_n\}$), given an image I:

$$\mathcal{L}^* = \arg\max_{\mathcal{L}}[P(\mathcal{L}|I)] \propto \arg\max_{\mathcal{L}}[P(I|\mathcal{L})\ P(\mathcal{L})] \tag{2}$$

In order to achieve efficient inference, most of the existing methods over-constraint the shape prior, $P(\mathcal{L})$. A very common strategy is to assume that the parts of the model form a tree-structure [10, 22, 31]. The tree property gives good results with relatively simpler object classes such as airplanes, motorcycles, and horses. However, it lacks the necessary loopy spatial constraints and produces unnatural deformations in the case of faces.

Statistical shape models have been used for "hard-detection" as well. These methods first generate "candidate sets" and *then* apply the shape constraints to eliminate the inconsistencies [26], and/or recover missing landmarks [2].

A fairly less explored approach for facial feature localization with statistical shape models is to use approximate inference. Sudderth et al. [25] demonstrated a method where they approximate the complex node potentials and

spatial relationships using Gaussian KDEs. In this work a very simple, hand-made, five-node graph is used to model the face. Nodes are defined to be high-dimensional feature vectors, representing *both* the location and the appearance of the corresponding part (e.g. left eye).

Our contributions in this paper are three-fold: 1) We propose a novel PGM-based framework, which addresses the limitations of existing facial feature localization methods, 2) Unlike other face models, the topology of our model is *learned* from training data, *not* manually set, 3) With greater flexibility in the allowed shape deformations, and still capturing the necessary loopy spatial constraints, the proposed approach advances the localization accuracy of the state-of-the-art in generic model fitting.

3. Proposed Approach

3.1. Background Information

3.1.1 Probabilistic Graphical Models (PGMs)

PGMs use a graph-based representation as the basis for compactly encoding complex joint distributions over multiple, high-dimensional random variables [18]. An undirected graph $\mathcal{G}$ is defined by a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$. Each node $s \in \mathcal{V}$ represents either an unobserved, or hidden, random variable x_s, or a noisy local observation y_s. Following the notation in [25], the neighborhood of a node $s \in \mathcal{V}$ is defined as $\Gamma(s) \triangleq \{t|(s,t) \in \mathcal{E}\}$.

In undirected, pairwise PGMs (see Figure 2) the joint distribution over all variables $p(x,y)$ factorizes as:

$$p(x,y) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \phi_{s,t}(x_s, x_t) \prod_{s\in\mathcal{V}} \phi_s(x_s, y_s) \quad (3)$$

where Z is a normalization constant, $\phi_{s,t}(x_s, x_t)$ is the *compatibility potential* between nodes s and t, and $\phi_s(x_s, y_s)$ is the *observation potential* of node s.

While the joint distribution $p(x,y)$ is hard to estimate, in many applications, the real interest is in the computation of conditional marginal distributions $p(x_s|y)$ for all $x_s \in \mathcal{V}$.

3.1.2 Belief Propagation (BP)

BP provides a convenient way for computing the conditional marginal distributions $p(x_s|y)$. At iteration n of the BP algorithm, each node $t \in \mathcal{V}$ sends a message $m^n_{t,s}(x_s)$ to each of its neighbors $s \in \Gamma(t)$:

$$m^n_{t,s}(x_s) = \alpha \int_{x_t} \phi_{s,t}(x_s, x_t)\phi_t(x_t, y_t) \prod_{u\in\Gamma(t)\setminus s} m^{n-1}_{u,t}(x_t)dx_t \quad (4)$$

where α denotes a proportionality constant.

At any iteration n, the *belief* of node s about the hidden variable x_s may be computed as follows:

$$\hat{p}^n(x_s|y) = \alpha\phi_s(x_s, y_s) \prod_{u\in\Gamma(s)} m^n_{u,s}(x_s) \quad (5)$$

BP algorithm guarantees that the node beliefs will converge to the correct conditional marginals in singly connected graphs [21]. Even though there is little theoretical analysis on the performance of BP in graphs with loops ([28, 29]), loopy BP has shown excellent empirical performance in a number of applications [1, 13].

3.1.3 Nonparametric Belief Propagation (NBP)

Equation 4 may be evaluated analytically only when both the compatibility and the observation potentials have special forms. When both are Gaussians, the calculations are straightforward since the product of a number of Gaussian densities is another Gaussian. When either potential is a Gaussian mixture and the other one is a Gaussian or a Gaussian mixture, still the integration is straightforward, but now the number of mixture components increase exponentially at every iteration. And when the potentials do not have special forms, analytical evaluation of the integral in Equation 4 becomes intractable.

In order to address this limitation of the BP algorithm, Sudderth et al. [25] and Isard [16] independently developed almost identical algorithms, which incorporate particle filters into the BP framework.

In these algorithms, nonparametric Gaussian KDEs [24] are used to represent the messages at each iteration. Then the BP update rule defined in Equation 4 becomes:

$$m^n_{t,s}(x_s) = \sum_{i=1}^{M} w^{(i)}_s \mathcal{N}(x_s; \mu^{(i)}_s, \Lambda_s) \quad (6)$$

where $w^{(i)}_s$ is the weight associated with the i^{th} kernel with mean $\mu^{(i)}_s$ and bandwidth Λ_s.

Given input messages $m_{u,t}(x_t)$ for each $u \in \Gamma(t)\setminus s$, the output message $m_{t,s}(x_s)$ is then computed as follows:

1. Draw M independent samples $\{\hat{x}^{(i)}_t\}^M_{i=1}$ from $\phi_t(x_t, y_t) \prod_{u\in\Gamma(t)\setminus s} m^{n-1}_{u,t}(x_t)dx_t$, and
2. For each $\{\hat{x}^{(i)}_t\}^M_{i=1}$ sample $\hat{x}^{(i)}_s \sim \phi_{s,t}(x_s, x_t = \hat{x}^{(i)}_s)$.

Details may be found in [25].

3.2. Shape Model

In our formulation, $x = \{x_s|s \in \mathcal{V}\}$ represent the 2D landmark locations, and $y = \{y_s|s \in \mathcal{V}\}$ represent the corresponding local image observations.

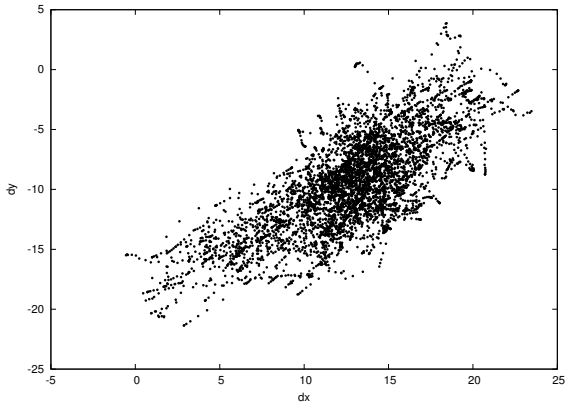


Figure 3. Example $(x_s - x_t)$ scatter plot. Note that an anisotropic Gaussian would model this distribution fairly well.

As illustrated in Figure 2(c), the proposed approach may be thought as modeling the prior shape knowledge in terms of *multiple*, *weak*, *pairwise* spatial relationships. In order to fully specify this shape model, one needs to define both the pairwise compatibility potentials *and* the topology of the underlying graph.

3.2.1 Compatibility Potentials

Anisotropic Gaussians are used to model the pairwise compatibility potentials:

$$\phi_{s,t}(x_s, x_t) = \mathcal{N}((x_s - x_t); \mu_{s,t}, \Sigma_{s,t}) \tag{7}$$

where $\mu_{s,t}$ is the mean, and $\Sigma_{s,t}$ is the covariance matrix of the Gaussian. Both of these parameters are learned from training data.

This potential encloses the prior shape knowledge between two landmarks, since given the location of a landmark x_t and the potential $\phi_{s,t}(x_s, x_t)$, one may estimate the likely locations of landmark x_s by:

$$\phi_{s,t}(x_s, x_t = \hat{x}_t) = \mathcal{N}(x_s; \mu_{s,t} + \hat{x}_t, \Sigma_{s,t}) \tag{8}$$

As Figure 3 illustrates, anisotropic Gaussians model the compatibility potentials well. Furthermore, one may estimate the "importance" of a particular pairwise potential within the model, simply by examining the learned covariance matrices. A potential with a smaller $\Sigma_{s,t}$ will have a higher precision, and hence would be more informative than a potential with a larger $\Sigma_{s,t}$.

3.2.2 Graph Topology

One of the primary advantages of using a PGM-based shape model is the flexibility in determining the graph topology. The possibilities range from a loose, singly connected graph, to a very strict, fully connected graph. A singly connected graph would reminiscence Snakes [17], whereas parametrized shape models [5] would be considered densely connected graphs.

In this work, the graph topology of the shape model is *learned* from training data. For each node s, the neighborhood $\Gamma(s)$ is determined using the computed compatibility potentials. Only the k most informative (smaller $\Sigma_{s,t}$) nodes are connected to node s. This approach allows for capturing a lot of shape knowledge in a fairly simple graph. Figure 4 illustrates the computed topologies for $k = 1, 2, 3, 4$.

Please note that the proposed approach actually generates a "class" of spatial models, rather than just a single one. Hence, the appropriate level of flexibility may be chosen with respect to the application.

3.3. Observation Model

A variety of observation models have been used in the facial feature localization literature. These methods vary from using gradients as in the case for Snakes [17], to using holistic error images in the case of AAMs [20]. CLMs, on the other hand, use "local experts" (i.e. local patch detectors), and have shown to perform superior [23].

The local experts used in this work are linear support vector machines (SVMs) [6] and the features used are 3×3 histograms of oriented-gradients (HOGs) [8] with $6-$bin histograms in each cell.

Inspired by the results of Saragih et al. in [23], the observation potentials are defined to be the nonparametric isotropic Gaussian KDEs [24] of the expert response maps:

$$\phi_s(x_s, y_s) = \sum_{z_i \in \Psi_s} \pi_{z_i} \mathcal{N}(x_s; z_i, \rho I) \tag{9}$$

where Ψ_s denotes the set of integer pixel locations within a square region centered at x_s, ρ is the bandwidth of the kernels, and π_{z_i} is the probabilistic expert response at location z_i.

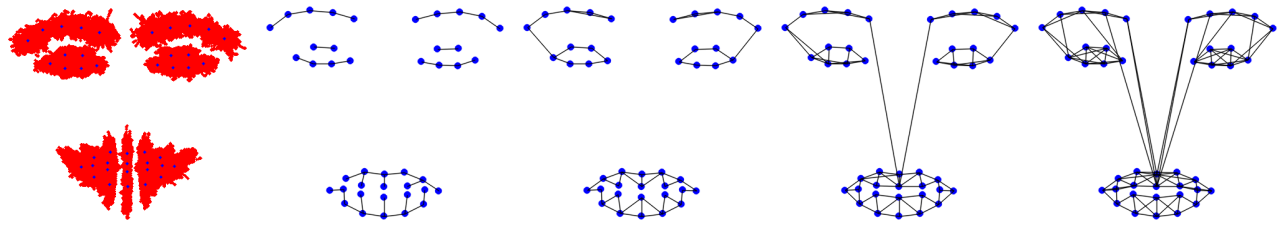
This observation potential has two advantages:

1. Its Gaussian mixture form fits well into the NBP framework (much better than than the one proposed in [25]), and allows for the employment of efficient sampling methods (e.g. [15]), and

2. It estimates the true response maps much better than the existing parametric methods [23].

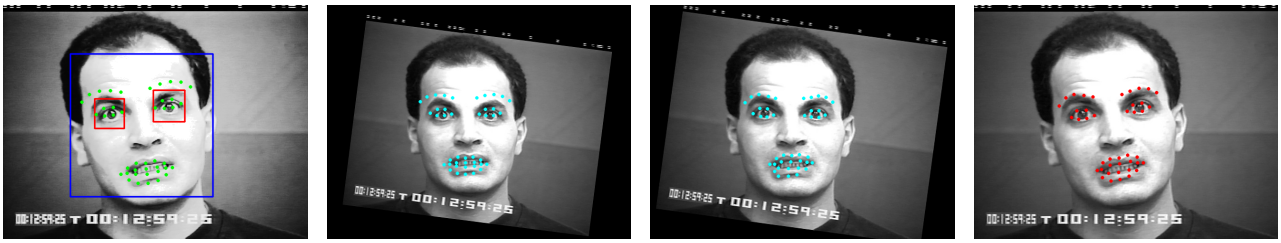
4. Experiments and Results

Extensive qualitative and quantitative experiments are performed on The Extended Cohn-Kanade Dataset (CK+) [19] and random images obtained from the Internet. These experiments contain subjects with different ethnicities, performing acted and/or spontaneous expressions. Imaging



(a) GPA. (b) $k = 1$. (c) $k = 2$. (d) $k = 3$. (e) $k = 4$.

Figure 4. Results of Generalized Procrustes Analysis (GPA) and learned graph topologies for $k = 1, 2, 3, 4$.



(a) Initialization. (b) Iteration 1. (c) Iteration 2. (d) Result.

Figure 5. The fitting process. At each iteration the image is "similarity normalized" by aligning the current shape with the mean shape (best viewed in color and high-resolution).

conditions and quality change significantly between the examples.

4.1. Shape and Expert Training

"Ground-truth" landmarks provided by the CK+ dataset are used for both shape and local expert training. Chin and nose region landmarks are ignored since these landmarks contribute much less information in most applications.

For the shape training, first the landmarks are shape normalized using Generalized Procrustes Analysis [9] (see Figure 4(a)). Then the compatibility potential parameters, $\mu_{s,t}$ and $\Sigma_{s,t}$, are computed. Finally the graph topology is determined using the computed covariance matrices as illustrated in Figures 4(b)-4(e).

Figure 4 demonstrates a major advantage of the proposed algorithm. By learning the graph topology from training data, we effectively obtain the smallest graph that would capture the most prior shape knowledge.

The resulting graph, in the case of faces, is a very intuitive one, where the parts of the face (e.g. eyes, mouth, etc.) are densely connected, while the parts themselves are loosely connected. Such a model will allow for a high variability between the locations of the parts, but at the same time enforce more strict constrains on how the parts themselves may deform. In these experiments we used $k = 3$ connectivity.

24×24 patches are used to train the local experts. For each landmark, positive examples are obtained from 1000 randomly selected images. Approximately 8000 negative examples are extracted from the remaining landmarks and other randomly selected images. LIBSVM [3] library is used for the SVM training.

4.2. Testing

Unless otherwise specified, all test images are *automatically* initialized. Local "search window", Ψ_s, of Equation 9 is set to be a 23×23 region centered around the current estimate. ρ is set to 1 and finally $M = 200$ particles are used for belief propagation.

Algorithm convergence is determined using the node beliefs. At each iteration the beliefs (i.e. landmark locations) are computed and when none of the landmarks move more than 1.5 pixels in radius, the algorithm is terminated.

4.3. Handling Similarity Transforms

Similarity transforms may be incorporated into this model either by scaling and rotating the compatibility potentials (Equation 7), or by keeping them constant, but instead aligning the current landmark estimates *and* the image with the mean scale and rotation at each iteration. We followed the second approach since it also implicitly solves the scale and rotation variance of the experts. This fitting process is illustrated in Figure 5.

4.4. Qualitative Results

Qualitative results on the CK+ dataset are presented in Figure 6. As the figure illustrates, the proposed algorithm performs equally well in a wide variety of examples, where

Figure 6. Qualitative results of the proposed PGM-based approach (best viewed in color and high-resolution, green: ground truth, red: results).

both the facial expressions *and* the facial attributes of the subject change significantly. This is primarily due to the higher level of shape flexibility provided by the model.

4.5. Qualitative and Quantitative Comparisons

The proposed approach is compared with two existing methods in Figures 7 and 8: 1) the "Tree-model" by Zhu and Ramanan, which has been proposed very recently, and 2) CLM by Saragih et al., which may be considered the current state-of-the-art in facial feature localization.

A total of 5876 images from 327 sequences have been tested. For every sequence, the first frame is automatically initialized. Every other frame in the sequence is initialized with the results of the previous frame.

As Figures 7 and 8 illustrate, both CLM and the proposed method significantly outperforms the "Tree-model". Even though similar tree-models perform well in other applications (such as part-based object classification), for facial feature localization they are too flexible, and hence allow unnatural deformations in the shape.

Out of 5876 tested images, the proposed approach achieved a lower average error in 3468 images (59.02%), CLM achieved a lower average error in 2296 images (39.07%) and the Tree-model achieved a lower average error in 112 images (1.91%). Corresponding error distributions are presented in Figure 7.

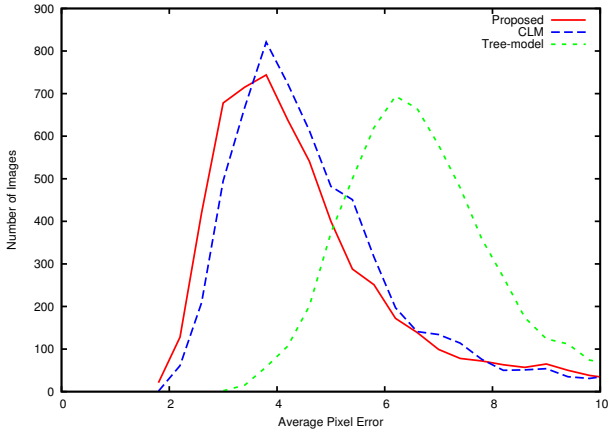
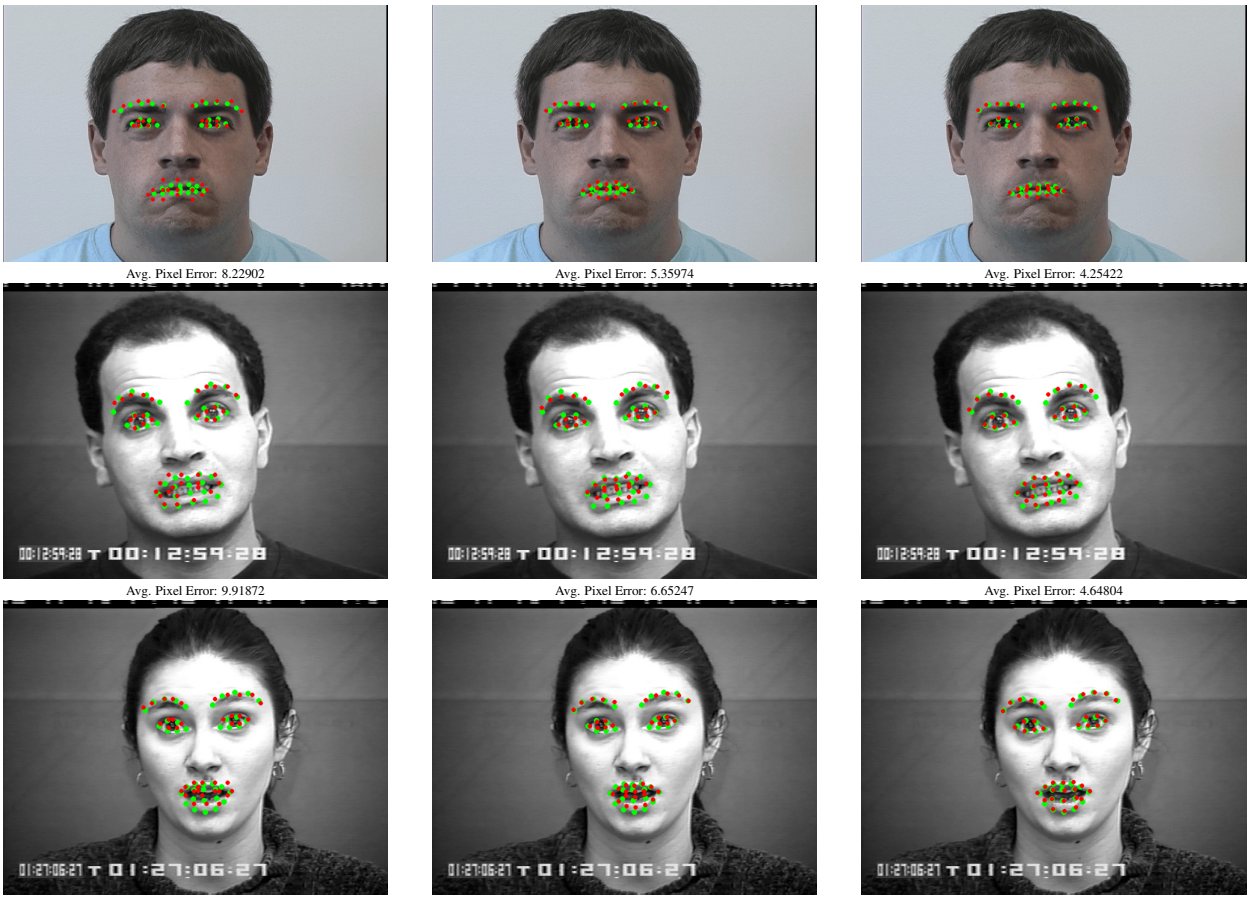


Figure 7. Quantitative comparison of Tree-model [31], CLM [23] and the proposed approach (best viewed in color).

4.6. Generalization

Even though the proposed algorithm is trained on a relatively small, fairly controlled dataset, as Figures 9 and 10 illustrate, it generalizes very well to real world images. This may be explained by two primary properties: 1) pairwise unimodal Gaussian compatibility potentials in the shape model allow for a great level of flexibility and generalization power, and 2) the HOG features capture the "generic" appearance properties of the landmarks very well.



(a) Tree-model [31]. (b) CLM [23]. (c) The proposed approach.

Figure 8. Qualitative comparison of Tree-model [31], CLM [23] and the proposed approach (best viewed in color and high-resolution, green: ground truth, red: algorithm-specific results, less green seen means a better fit).

4.7. Implicit Occlusion Handling

Unlike AAM and CLM, our proposed approach models the prior shape knowledge as pairwise local spatial relationships. Figure 10 illustrates an important advantage of this local model over the holistic approaches. Even with highly occluded faces: 1) the visible landmarks are not affected from the occlusion, and 2) reasonable predictions can be made about the occluded landmarks. Please note that the results in the figure are obtained without *any* explicit occlusion handling mechanism.

5. Conclusion

In this paper, we proposed a novel approach for modeling the prior shape knowledge in the facial feature localization problem. Our framework is strict enough to capture all the necessary loopy spatial constraints on a face, yet flexible enough to generalize well to unseen expression and/or facial attributes.

Through extensive qualitative and quantitative results we showed that the proposed algorithm outperforms the state-of-the-art in terms of localization accuracy.

Our future work consists of: 1) expanding the proposed model to 3D, and 2) incorporating external applications into the PGM framework.

Acknowledgments: We would like to thank the anonymous reviewers for their valuable feedback.

References

[1] C. Berrou, A. Glavieux, and P. Thitimajshima. Near shannon limit error-correcting coding and decoding: Turbo-codes. In *Communications*, volume 2, pages 1064 –1070 vol.2, 1993.

[2] O. Celiktutan, H. Akakin, and B. Sankur. Multi-attribute robust facial feature localization. In *FG*, pages 1–6, 2008.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[4] T. F. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *ECCV*, volume 1407, pages 484–498. 1998.

Figure 9. Generalization to random Internet images (best viewed in color and high-resolution).

Figure 10. Implicit occlusion handling example (best viewed in color and high-resolution).

[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models their training and application. *Comput. Vis. Image Underst.*, 61:38–59, January 1995.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[7] D. Cristinacce and T. Cootes. A comparison of shape constrained facial feature detectors. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE Int. Conference on*, pages 375 – 380, may 2004.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. CVPR, 2005.

[9] I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. John Wiley & Sons, 1998.

[10] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *Int. J. of Comp. Vis.*, 61(1):55–79, 2005.

[11] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages II–264–II–271 vol.2, 2003.

[12] M. A. Fischler and R. Elschlager. The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, C-22(1):67–92, 1973.

[13] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vis. *Int. J. of Comp. Vis.*, 40:25–47, 2000.

[14] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *ECCV*, volume 5302, pages 413–426. 2008.

[15] E. T. Ihler, E. B. Sudderth, W. T. Freeman, and A. S. Willsky. Efficient multiscale sampling from products of gaussian mixtures. In *In NIPS 17*. MIT Press, 2003.

[16] M. Isard. Pampas: real-valued graphical models for computer vision. In *CVPR*, 2003.

[17] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int. J. of Comp. Vis.*, 1:321–331, 1988.

[18] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[19] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, pages 94 –101, june 2010.

[20] I. Matthews and S. Baker. Active appearance models revisited. *Int. J. of Comp. Vis.*, 60:135–164, 2004.

[21] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[22] D. Ramanan and C. Sminchisescu. Training deformable models for localization. In *CVPR*, 2006.

[23] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shift. *Int. J. of Comp. Vis.*, 91:200–215, 2011.

[24] B. Silverman. *Density estimation for statistics and data analysis*. 1986.

[25] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. In *CVPR*, volume 1, pages 605–612, 2003.

[26] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, pages 2729–2736, 2010.

[27] Y. Wang, S. Lucey, and J. Cohn. Enforcing convexity for improved alignment with constrained local models. In *CVPR*, pages 1 –8, june 2008.

[28] Y. Weiss. Correctness of belief propagation in graphical models with loops. *Neural Comp.*, 12(1):1–41, 2000.

[29] Y. Weiss and W. T. Freeman. Correctness of belief propagation in gaussian graphical models of arbitrary topology. *Neural Computation*, 13(10):2173–2200, 2001.

[30] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. Von der Malsburg. Face recognition by elastic bunch graph matching. *PAMI*, 19(7):775–779, 1997.

[31] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.