# Active Conditional Models

Ying Chen[1] and Fernando De la Torre[2]

[1]School of IoT Engineering, Jiangnan University, China. e-mail: chenying@jiangnan.edu.cn

[2] Robotics Institute, Carnegie Mellon University. Pittsburgh, PA 15213, USA. e-mail: ftorre@cs.cmu.edu

***Abstract*— Matching images with large geometric and iconic changes (e.g. faces under different poses and facial expressions) is an open research problem in computer vision. There are two fundamental approaches to solve the correspondence problem in images: Feature-based matching and model-based matching. Feature-based matching relies on the assumption that features are stable across view-points and iconic changes, and it uses some unary, pair-wise or higher-order constraints as a measure of correspondence. On the other hand, model-based approaches such as Active Shape Models (ASMs) align appearance features with respect to a model. The model is learned from hand-labeled samples. However, model-based approaches typically suffer from lack of generalization to untrained situations.**

**This paper proposes Active Conditional Models (ACM) that combines the benefits of both approaches. ACM learns the conditional relation (both in shape and appearance) between a reference view of the object and other view-points or iconic changes. The ACM model generalizes better to untrained situations, because it has less number of parameters (less prone to overfitting) and directly learns variations w.r.t a reference image (similar to feature-based methods). Several examples in the context of facial feature matching across pose and expression illustrate the benefits of ACMs.**

## I. Introduction

Establishing correspondence between images is a fundamental problem in computer vision. Correspondence between images is needed in many computer vision algorithms such as object recognition [1], image registration [2], face detection and tracking [3], face recognition [5], and 3D reconstruction [4]. Approaches to solve for correspondence could be broadly divided in feature-based and model-based methods. Feature-based methods extract features in both images and use some unary [6], [18], pair-wise [7] or higher-order constraint [10] to solve correspondence. However, feature-based approaches typically fail when there is a large change in camera motion, pose or large iconic changes (e.g. eyes closed). As shown in Fig. 1(a), feature based approaches fail to detect correspondence when there are strong changes in pose and expression. On the other hand, model-based approaches incorporate a priori information about iconic and view-point changes. Many methods, such as Snakes [12], Parameterized Appearance Models (e.g Active Shape Models(ASM) [13], Active Appearance Models(AAM) [14], [27], 3D Morphable Models (3DMM) [15]) have been successfully applied to solve facial correspondence. However, these models typically suffer from lack of generalization to untrained samples.

Dr. De la Torre was partially supported by National Institute of Health Grant R01 MH 051435. Dr. Ying was supported by the China Scholarship Council Grant 2008110195.

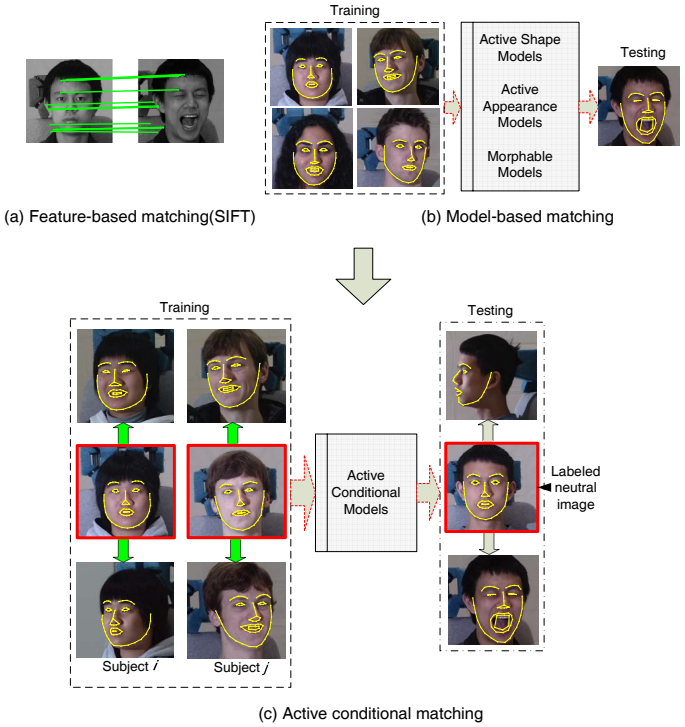



Fig. 1. (a) Feature-based (SIFT) method for matching two images. (b) Model-based approaches to solve the correspondence between image features and a model. (c) Active Conditional Models (ACM) learn a mapping from a reference image (red box) to other images with changes in pose and expression.

In this paper we propose a new statistical model, Active Conditional Models (ACMs), that benefits from both model-based and feature-based matching. ACMs learn the shape and appearance representation of one class of objects conditional to a reference image. ACMs outperform PAMs in generalization ability, because they learn the conditional information relative to a reference image. Unlike feature-based correspondence, ACMs constrain with a shape and appearance model. Unlike model-based methods, ACMs incorporates a reference image that reduces the amount of person-specific variability. Fig. 1 illustrates the differences between the three approaches to image matching.

The remainder of this paper is structured as follows: Section 2 provides a brief overview of related work. Section 3 gives details of learning and inference in ACMs. In Section 4, ACMs are extended by adding appearance. The experimental results are presented and discussed in Section 5.

## II. Previous Work

This section reviews previous work on feature-based and model-based image matching.

### A. Feature-based image matching

Image matching has been a central research topic in computer vision over the last few decades. Typical approaches to correspondence involve matching feature points between images. Lowe's SIFT descriptor [18] is one of the state-of-the-art methods to construct geometric invariant features to match rigid objects. SIFT and its extensions [24], [19], [25], [6], [23] have been successfully applied to many problems.

Alternatively, the correspondence problem can be formulated into a graph matching problem considering each feature point in the image as a node in a graph. Leordeanu and Herbert [7] proposed a spectral graph matching optimization algorithm using general unary and pairwise constraints. They built an affinity matrix between pair-wise points, and the correspondence is found by thresholding the leading eigenvectors. The problem of hyper-graph matching has been further studied [20], [10], considering higher-order interactions between tuples of features beyond pairwise.

Recently, researchers paid more attention to learn the optimal set of parameters for graph matching. Caetano *et al*. [11] made use of structural models improving the graph matching solution. Torresani *et al*. [21] proposed an energy minimization approach to establish correspondences for non-rigid motion, in which the parameter of the error function are learned by NIO algorithm [22]. Leordeanu and Herbert [9] introduced a unsupervised learning strategy to learn the weights in spectral matching.

### B. Model-based image matching

Model-based methods are able to solve the correspondence in difficult situations because the model encodes prior knowledge of the expected shape and appearance of an object class. ASMs [13] have been proven an effective method to model the shape and appearance of objects. ASMs build a model of shape variation by performing PCA on a set of landmarks (after Procrustes). For each landmark, the Mahalanobis distance between the sample and mean texture is used to assess the quality of the fit of a new shape. The fitting process is performed using a local search along the normal of the landmark. Later, the new positions are projected onto rigid and non-rigid basis. These two steps are alternated until convergence. Further extensions of ASMs include 3D Morphable Models [15], AAM [14] and kernel generalizations [27] among others. The work that is closest to the proposed one is the one done by Asthana *et al*. [16]. This approach [16] utilizes the MPEG-4 based facial animation system to generate virtual images having different poses. The set of virtual images is used as training data for a view-based AAM, and later locate landmarks in previously untrained face images.

## III. Active Conditional Shape Models (ACSMs)

This section describes the details of Active Conditional Shape Models (ACSMs). Unlike ASMs, ACSMs learn a mapping between a neutral shape and shape variations of the same individual under different poses and expressions.

### A. ACSMs Energy Function

Let $\boldsymbol{F}_i \in \Re^{3\times d_f}$ (see footnote for notation[1]) be the homogeneous representation for $d_f$ landmarks located in the neutral (or reference) view of the $i^{th}$ subject. Fig. 1.c shows two examples of labeled face images with 66 landmarks under different expressions and poses. $\boldsymbol{F}_i$ has the following structure:

$$\boldsymbol{F}_i = \begin{pmatrix} u_{i1} & u_{i2} & \dots & u_{id_f} \\ v_{i1} & v_{i2} & \dots & v_{id_f} \\ 1 & 1 & \dots & 1 \end{pmatrix} \quad (1)$$

where $(u_{ij}, v_{ij})^T$ denotes $j^{th}$ landmark coordinates for the $i^{th}$ subject.

ACSMs learn a mapping between the neutral (or reference) view of the object and the object under different rigid and non-rigid transformations (see Fig. 1(c)). Let $\boldsymbol{q}_i^b \in \Re^{2d_g\times 1}$ denote the $i^{th}$ subject under $q$ different configurations, and $d_g$ denotes the number of landmarks. $i = 1,2,...,n$ indexes the person identity, $b = 1,2,...,p$ denotes possible deformations (e.g. viewpoint or expression changes) and $n$ is number of subjects . $\boldsymbol{q}_i^b$ has the following structure:

$$\boldsymbol{q}_i^b = \begin{pmatrix} x_{i1}^b & y_{i1}^b & \dots & x_{id_g}^b & y_{id_g}^b \end{pmatrix}^T \quad (2)$$

where $(x_{ij}^b, y_{ij}^b), j = 1,2,...d_g (d_g \leq d_f)$ is the coordinate of $j^{th}$ point in the $b^{th}$ deformed shape of the $i^{th}$ individual.

The ACSM minimizes:

$$E_{ACSM}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}) = \sum_{i=1}^{n}\sum_{b=1}^{p} \|\boldsymbol{q}_i^b - (\sum_{j=1}^{k} c_{ij}^b \boldsymbol{B}_j) vec(\boldsymbol{A}_i^b \boldsymbol{F}_i)\|_2^2, \quad (3)$$

where $\boldsymbol{A}_i^b \in \Re^{2\times 3}$ is an affine matrix that compensates for geometric changes (e.g. rotation, scale). $\boldsymbol{B}_j \in \Re^{2d_g\times 2d_f}$ is a basis such that a linear combination weighted by the coefficient $c_{ij}^b$ models non-rigid deformations and 3D rigid transformations (that an affine transformation cannot recover). Fig. 2 illustrates the deformation process from a neutral face $\boldsymbol{F}_i$ to faces under different view point or non-rigid deformations due to changes in expressions and pose.

[1]Bold capital letters denote matrices $\boldsymbol{D}$, bold lower-case letters a column vector $\boldsymbol{d}$. $\boldsymbol{d}_j$ represents the $j^{th}$ column of the matrix $\boldsymbol{D}$. All non-bold letters represent scalar variables. $d_{ij}$ denotes the scalar in the row $i$ and column $j$ of the matrix $\boldsymbol{D}$ and the scalar $i$-th element of a column vector $\boldsymbol{d}_j$. *diag* is an sperator that transforms a vector to a diagonal matrix or takes the diagonal of the matrix into a vector. $\mathbf{1}_k \in \Re^{k\times 1}$ is a vector of ones. $||\boldsymbol{d}||_2^2$ denotes the norm of the vector $\boldsymbol{d}$. $\circ$ denotes the Hadamard or point-wise product, $\otimes$ denotes the Kronecker product, $vec(\cdot)$ represents the vec operator that converts a matrix into a column vector, and $mod(x,y)$ denotes the modulus after $x$ dividing by $y$.

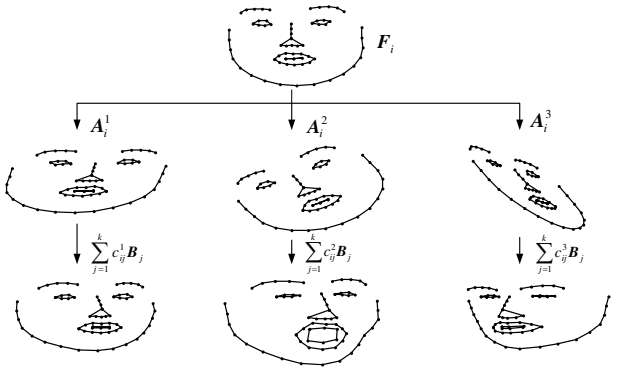


Fig. 2. Shape deformation process

Equation 3 can be re-written as:

$$E_{ACSM}(\boldsymbol{U},\boldsymbol{B},\boldsymbol{C}) = \sum_{l=1}^{m} \|\boldsymbol{q}_l - (\sum_{j=1}^{k} c_{jl}\boldsymbol{B}_j)vec(\boldsymbol{A}_l\boldsymbol{F}_l)\|_2^2 = \sum_{l=1}^{m} \|\boldsymbol{q}_l - (\sum_{j=1}^{k} c_{jl}\boldsymbol{B}_j)\boldsymbol{u}_l\|_2^2, \quad (4)$$

where $m = np$, $l$ is an index that includes all possible instances of a subject $i$ and all instances of possible configurations $b$, satisfying $i = \lceil l/p \rceil$ and $b = mod(l,p)$, and $\boldsymbol{u}_l = vec(\boldsymbol{A}_l\boldsymbol{F}_l) \in \Re^{2d_f \times 1}$.

### B. Optimization

To learn the ACSMs' parameters we have to minimize Eq. 4 with respect to the non-rigid transformation matrix $\boldsymbol{B}$, the local coefficients $\boldsymbol{c}_l$ for every subject and every instance of deformation, and the affine transformation $\boldsymbol{A}_l$. We use an alternated least square (ALS) strategy to monotonically reduce the error in each step. ALS alternates between optimizing for $\boldsymbol{B}$ while $\boldsymbol{A}_l$ and $\boldsymbol{c}_l$ are fixed and optimizing $\boldsymbol{A}_l$ when $\boldsymbol{c}_l$ and $\boldsymbol{B}$ are fixed. The minimization scheme is shown in the table of Algorithm 1, where:

$$\boldsymbol{Q} = [\boldsymbol{q}_1 \quad \boldsymbol{q}_2 \quad \dots \quad \boldsymbol{q}_m], \qquad \boldsymbol{B} = [\boldsymbol{B}_1 \quad \boldsymbol{B}_2 \quad \dots \quad \boldsymbol{B}_k]$$
$$\boldsymbol{U} = [\mathbf{1}_{\boldsymbol{k}} \otimes \boldsymbol{u}_1 \quad \mathbf{1}_{\boldsymbol{k}} \otimes \boldsymbol{u}_2 \quad \dots \quad \mathbf{1}_{\boldsymbol{k}} \otimes \boldsymbol{u}_m],$$
$$\boldsymbol{C} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1m} \\ c_{21} & c_{22} & \dots & c_{2m} \\ \dots & \dots & \dots & \dots \\ c_{k1} & c_{k2} & \dots & c_{km} \end{pmatrix} \quad (5)$$

## IV. Active Conditional Appearance Models (ACAMs)

Similar to ACSMs, ACAMs learn the conditional relation between a reference (or neutral) appearance and the appearance of the same subject under different conditions. Similar to feature-based methods, it uses a reference image.

---

**Algorithm 1** Learning Active Conditional Shape Models

1. Initialize parameters $\boldsymbol{B}$ and $\boldsymbol{C}$.

2. For each instance $l$, minimize Eq. 4 over $\boldsymbol{A}_l$ for fixed $\boldsymbol{B}$ and $\boldsymbol{c}_l$

$$\boldsymbol{A}_l^* = (\sum_{j=1}^{k} c_{jl}\boldsymbol{B}_j)^{-1}\boldsymbol{q}_l\boldsymbol{F}_l^T(\boldsymbol{F}_l\boldsymbol{F}_l^T)^{-1} \quad (6)$$

3. For each instance $l$, minimize Eq. 4 over $\boldsymbol{c}_l$ for fixed $\boldsymbol{A}_l(\boldsymbol{u}_l)$ and $\boldsymbol{B}$

$$\boldsymbol{c}_l^* = (\boldsymbol{Z}_1^T\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1^T\boldsymbol{q}_l \quad (7)$$

where $\boldsymbol{Z}_1 = \boldsymbol{BP}$, and $\boldsymbol{P} = diag(\mathbf{1}_k) \otimes \boldsymbol{u}_l$

4. Minimize Eq. 4 over $\boldsymbol{B}$ for fixed $\boldsymbol{A}(\boldsymbol{U})$ and $\boldsymbol{C}$

$$\boldsymbol{B}^* = \boldsymbol{Q}\boldsymbol{Z}_2^T(\boldsymbol{Z}_2\boldsymbol{Z}_2^T)^{-1} \quad (8)$$

where $\boldsymbol{Z}_2 = (\boldsymbol{C} \otimes \mathbf{1}_{2d_f}) \circ \boldsymbol{U}$.

5. If not converged, return to step 2.

---

ACAMs are similar to ACSMs but no affine transformation matrix $\boldsymbol{A}_l$ is necessary. We use $l$ to index all possible instances of a subject $i$ under all possible configurations $b$. $d$ is the dimension of the appearance features. Let $\boldsymbol{x}_l \in \Re^{d\times 1}$ be the neutral appearance of the $l^{th}$ instance, and $\boldsymbol{y}_l \in \Re^{d\times 1}$ be the deformed appearance of the $l^{th}$ instance, then the energy function is:

$$E_{ACAM}(\boldsymbol{B},\boldsymbol{C}) = \sum_{l=1}^{m} \|\boldsymbol{y}_l - (\sum_{j=1}^{k} c_{jl}\boldsymbol{B}_j)\boldsymbol{x}_l\|_2^2 \quad (9)$$

where $m$ is the number of appearance samples with all possible deformations. The ALS algorithm to fit ACAMs is illustrated in the table for Algorithm 2.

---

**Algorithm 2** Learning Active Conditional Appearance Model

1. Initialize parameter $\boldsymbol{B}$.

2. For each instance $l$, minimize Eq. 9 over $\boldsymbol{c}_l$ for fixed $\boldsymbol{B}$.

$$\boldsymbol{c}_l^* = (\boldsymbol{Z}_1^T\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1^T\boldsymbol{Y}_l \quad (10)$$

where $\boldsymbol{Z}_1 = \boldsymbol{BP}$, $\boldsymbol{Y} = [\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \dots \quad \boldsymbol{y}_m]$, and $\boldsymbol{P} = diag(\mathbf{1}_k) \otimes \boldsymbol{x}_l$.

3. Minimize Eq. 9 over $\boldsymbol{B}$ for fixed $\boldsymbol{C}$.

$$\boldsymbol{B}^* = \boldsymbol{Y}\boldsymbol{Z}_2^T(\boldsymbol{Z}_2\boldsymbol{Z}_2^T)^{-1} \quad (11)$$

where$\boldsymbol{Z_2} = (\boldsymbol{C} \otimes \mathbf{1}_{2d_f}) \circ \boldsymbol{X}$, and $\boldsymbol{X} = [\mathbf{1}_{\boldsymbol{k}} \otimes \boldsymbol{x}_1 \quad \mathbf{1}_{\boldsymbol{k}} \otimes \boldsymbol{x}_2 \quad \dots \mathbf{1}_{\boldsymbol{k}} \otimes \boldsymbol{x}_m]$.

4. If not converged, return to step 2.

---

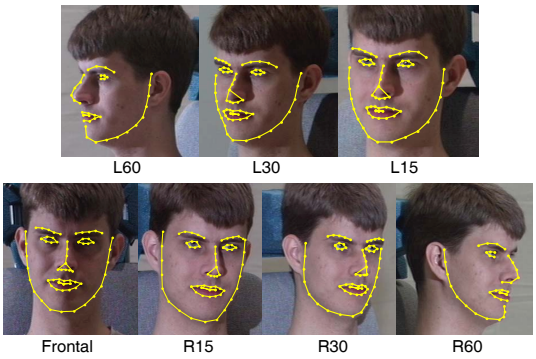



Fig. 3. Manual labels for one subject with changes in pose.

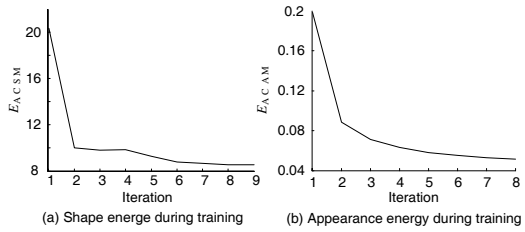



Fig. 4. Training error for ACSMs and ACAMs.

## V. Experiments and Discussion

We evaluated the proposed ACMs on the problem of facial feature detection across pose and expression using the CMU Multi-PIE face dataset [17]. The CMU Multi-PIE database has a total of 337 subjects, with 129 subjects present in all four sessions, under 15 different viewpoints, 6 expressions and 19 illuminations. 108,566 images under different illuminations with different subject ID, viewpoints and expressions are manually labeled.

Section V-A describes the training procedure for ACMs. Section V-B compares the generalization performance of ACMs against ASMs using the same training data. Section V-C provides results of the fitting error.

### A. Training Active Conditional Models

In all experiments, we used 8,240 labeled sample pairs (each pair contains one frontal image and one image under different pose or expression) from the CMU Multi-PIE database. We used 360 pairs for testing and the rest for training and cross-validation. As shown in Fig. 3, 66 salient points were manually labeled for the pose ranging from $-45$ to 45 degrees, and 38 points in the range from $-90$ to $-45$ and 45 to 90 degrees.

We used 3,729 sample pairs for training the ACSM (Algorithm 1) and the ACAM (Algorithm 2). Fig. 4 shows the training error for ACSMs and ACAMs. As expected, the error decreases monotonically with the number of iterations. Fig. 5 shows the synthesis of different shapes for different $\boldsymbol{A}_l$ and $\boldsymbol{c}_l$ parameters. It is worth pointing out that unlike ASMs, a linear ACSM can represent strong changes in pose in a continuous manner (see Fig. 5).

In our experiments, the dimension of $\boldsymbol{B}$, *i.e.* parameter $k$ in ACMs, is determined with 10-fold cross-validation in which the 4,150 validation sample pairs are partitioned into

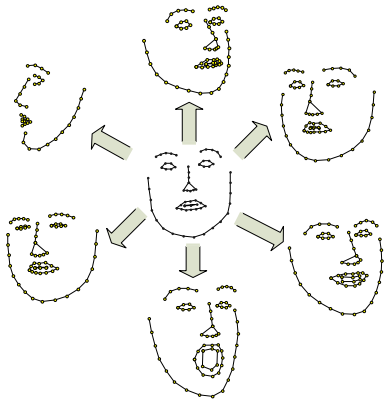


Fig. 5. Synthesized shapes using ACSMs.

10 subsamples. In the cross-validation set the optimal $k$ was estimated to be $k = 21$ for ACSM, and $k = 6$ for AASM. It is worth pointing out, that to make a fair comparison ASMs were trained with exactly the same data as ACMs (including the neutral image). In the case of ASMs, the dimension of shape and appearance is 30 and 58 respectively, which preserves 95% of PCA energy.

### B. Generalization of ACMs

This section analyzes the generalization properties of ACSMs and compares it to ASMs. In this section, we assume that the landmarks for the testing images (not training subjects) are known.

We took 49 untrained subjects (360 sample pairs) for testing and compared the shape reconstruction performance of ASM [14] and ACMs. That is, given the labeled landmarks in the testing sample $l$, we computed the shape reconstruction error as follows:

$$e_l^s = \frac{1}{d_g}\|\boldsymbol{g}_l^s - \boldsymbol{r}_l^s\|, \tag{12}$$

where $\boldsymbol{g}_l^s$ be the ground-truth shape vector, and $\boldsymbol{r}_l^s = (\sum_{j=1}^k c_{jl}\boldsymbol{B}_j) vec(\boldsymbol{A}_l\boldsymbol{F}_l)$ is the shape vector in the case of the ACSM (after converge). $\boldsymbol{F}_l$ is the shape vector of $l^{th}$ frontal sample, $\boldsymbol{B}$ is the learned basis, $\boldsymbol{A}_l$ and $\boldsymbol{c}_l$ are computed iteratively using Eq. (6) and Eq. (7) with $\boldsymbol{q}_l = \boldsymbol{g}_l^s$.

Fig. 6 shows the shape reconstruction error for the 360 samples. We display the reconstruction error and the histogram of the error for the ACSM and ASM. As can be observed, for most samples the shape error (assuming the landmarks are known) for ACSMs is lower than for ASMs.

Similarly, to evaluate the generalization of ACAMs to untrained samples, the appearance similarity error $e_{p,l}^a$ for the $p^{th}$ salient point $(p = 1,2,...,d_g)$ in the $l^{th}$ testing sample is defined as:

$$e_{p,l}^a = \frac{E_{i\in N_1(p)}\{\|\boldsymbol{g}_{l,p,i}^a - \boldsymbol{r}_{l,p,i}^a\|\}}{E_{j\in\{N_2(p)-N_1(p)\}}\{\|\boldsymbol{g}_{l,p,j}^a - \boldsymbol{r}_{l,p,j}^a\|\}}, \tag{13}$$

where $\boldsymbol{g}_{l,p,i}^a$ denotes the ground-truth appearance representation for the $i^{th}$ neighbor of point $p$ in $l^{th}$ sample, $\boldsymbol{r}_{l,p,i}^a$ the synthesized one, $E\{\cdot\}$ is the mean operator, and $N_1(p)$ and

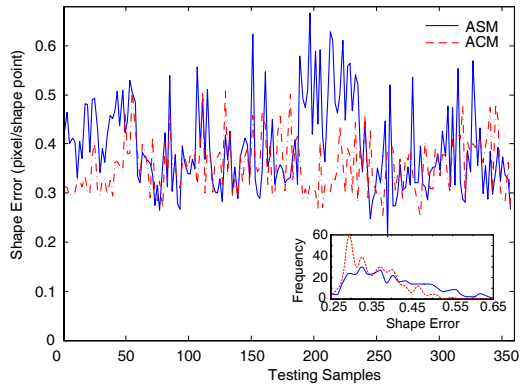



Fig. 6. Comparison of shape reconstruction error.

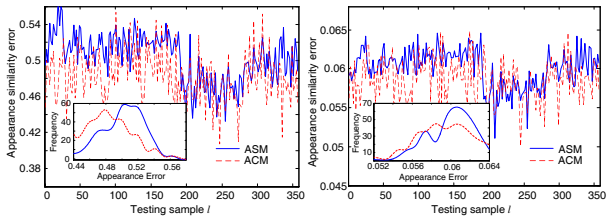



Fig. 7. Appearance error. (a) Size of $N_1(p)$ equals $5\times5$, and size of $N_2(p)$ equals $17\times17$; (b) (a) Size of $N_1(p)$ equals $1\times1$, and size of $N_2(p)$ equals $9\times9$

$N_2(p)$ represent small-sized and large-sized neighbors of $p$. $\|\boldsymbol{g}^a-\boldsymbol{r}^a\|$ is the appearance error between the image patch and its reconstruction by the ACAM.

Smaller values $e^a_{p,l}$ indicates that for each point $p$, the appearance difference of its close neighbor is smaller than the differences of its distant neighbors, which means that the fitting/searching process is more likely to find the right position of $p$. The appearance similarity error for the sample $l$ is the average of the appearance similarity error for each $p$, that is denoted by $e^a_l$.

Fig. 7 shows two graphics with the error for two different sizes of $N_1(p)$ and $N_2(p)$. In the first graphic $N_1(p)$ has the size of $5\times5$, and $N_2(p)$ has the size of $17\times17$. In the second graphic, $N_1(p)$ has the size of $1\times1$, and $N_2(p)$ has the size of $9\times9$. It is clear from Fig. 7 that independently of the neighbor size, ACAMs have a smaller appearance similarity error. This typically translates that the the ACAM is more likely to have a local minima in the expected position of the landmark.

### C. Fitting Error in ACMs

Previous section has shown the generalization properties of ACMs assuming the landmarks in the testing images are known. This section compares the fitting performance, when the landmarks in the testing images are unknown.

The ACMs search for landmarks is similar to the ASM search. First, we use a face detector to locate the face, and initialize the ACM with a mean shape. In each iteration and for each landmark point, the ACM's search is done along the normal of the shape profile and it selects the location with smaller appearance error. The new location is projected into the ACSMs. This process is repeated until the difference

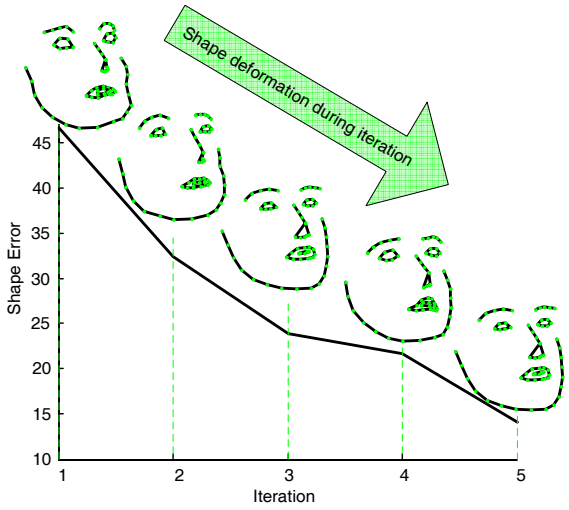



Fig. 8. Shape deformation during the search process.

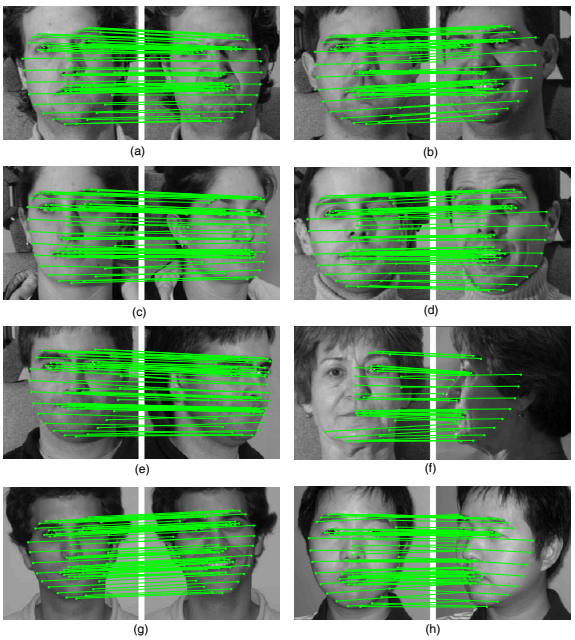
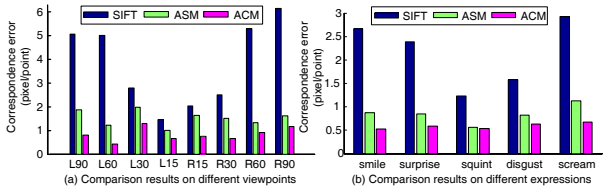



Fig. 9. Correspondence results with different viewpoints and expressions.

between the current synthesized shape and image points is less than an average of 0.3 pixel/point. We use the same graylevel representation to build a model for ASMs and ACMs.

Fig. 9 shows ACMs matching results across pose and expression, given the landmarks in the neutral image. Images in Fig. 9(a)-(h) are from Multi-PIE, and images in Fig. 9(g)-(h) are images taken with a regular camera in an unstructured environment. As can be seen, ACMs are able to successfully match across pose and expression. Fig. 8 illustrates how the shape varies with each iteration when fitting an untrained image.

Fig. 10 illustrates the correspondence error between facial features under different viewpoints (from -90 to 90) and different expressions (e.g., smile, disgust). We compared



(a) Comparison results on different viewpoints (b) Comparison results on different expressions

Fig. 10. Matching accuracy comparison. (a) images undergoing different viewpoints; (b) images undergoing different expressions.

three approaches: feature-based correspondence, ASMs and ACMs. For feature matching, we used SIFT as the feature to get sparse feature correspondence and interpolate them to get the landmarks in the expected locations. The correspondence error is defined as $e_l = \frac{1}{d_g}\|\boldsymbol{g}_l - \boldsymbol{r}_l\|$, where subscript $l$ denotes different deformation instances, such as smile, surprise or viewpoint changes. $\boldsymbol{r}_l$ is the ground truth landmarks, and $\boldsymbol{g}_l$ is the reconstructed shape (vectorized) from the neutral image.

Fig. 10 shows a quantitative evaluation of the shape error (difference between the ground truth labels and the landmarks provided by the algorithm). SIFT-based correspondence performs badly under strong changes in pose or expression. As expected, ACMs outperform ASMs because it is less prone to overfitting and learns a conditional relation with a reference image.

## VI. Conclusion

This paper proposes ACMs, a model that learns conditional relations between images of an object under different configurations and a reference image. ACMs learn a discriminative bilinear shape and appearance model. Experimental results on the CMU Multi-pie face database show that ACMs outperforms ASMs and feature-based methods (using SIFT) when matching face images undergoing changes in pose and expression. Further research needs to be addressed to avoid local minima in the fitting process.

## VII. Acknowledgments

Dr. De la Torre was partially supported by National Institute of Health Grant R01 MH 051435. Dr. Ying was supported by the China Scholarship Council Grant 2008110195.

## References

[1] M. F. Demirci, A. Shokoufandeh, Y. Keselman, L. Bretzner and S. Dickinson, "Object Recognition as Many-to-Many Feature Matching", *International Journal of Computer Vision*, vol. 69, 2006, pp. 203–222.

[2] D. Shen, "Fast Image Registration by Hierarchical Soft Correspondence Detection", *Pattern Recognition*, vol. 42, 2009, pp. 954–961.

[3] D. Boley, R. Maier, W. Zhang, Q. Wang and X. Tang, Real Time Feature Based 3-D Deformable Face Tracking", *in Proceedings of the 10th European Conference on Computer Vision*, 2008, pp. 720–732.

[4] S. Agarwal and N. Snavely and I. Simon and S. M. Seitz and R. Szeliski, "Building Rome in a Day", *in IEEE International Conference on Computer Vision*, 2009, pp. 72–79.

[5] D. R. Kisku, M. Tistarelli, J. K. Sing and P. Gupta, "Face Recognition by Fusion of Local and Global Matching Scores using DS Theory: An Evaluation with Uni-classifier and Multi-classifier Paradigm", *in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 60–65.

[6] J. M. Morel and G. Yu, "ASIFT-A New Framework for Fully Affine Invariant Image Comparison", *SIAM Journal on Imaging Sciences*, vol. 2, 2009, pp. 438–469.

[7] M. Leordeanu and M. Hebert, "A Spectral Technique for Correspondence Problems using Pairwise Constraints", *in International Conf. of Computer Vision*, 2009, pp. 1482–1489.

[8] M. Leordeanu, M. Hebert and R. Sukthankar, "Beyond Local Appearance: Category Recognition from Pairwise Interactions of Simple Features", *in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[9] M. Leordeanu and M. Hebert, "Unsupervised Learning for Graph Matching", *in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 864–871.

[10] O. Duchennel, F. Bach, I. Kweon and J. Ponce, "A Tensor-Based Algorithm for High-Order Graph Matching", *in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 1980–1987.

[11] T. S. Caetano, J. J. McAuley, L. Cheng, Q. V. Le and A. J. Smola, "Learning Graph Matching", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, vol. 31, pp. 1048–1058.

[12] M. Kass, A. Witkin, and D. Terzopoulos, Snakes - Active Contour Models, *International Journal of Computer Vision*, vol. 1, 1987, pp 321-331.

[13] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, "Active shape models: Their Training and Application", *Computer Vision and Image Understanding*, vol. 61, 1995, pp. 38–59.

[14] T. F. Cootes, G. J. Edwards and C. J. Taylor, "Active Appearance Models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, 2001, pp. 681–685.

[15] V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3d Faces", *in Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[16] A. Asthana, R. Goecke, N. Quadrianto and T. Gedeon, "Learning Based Automatic Face Annotation for Arbitrary Poses and Expressions from Frontal Images Only", *in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 1635–1642.

[17] R. Gross, I. Matthews, J. Cohn, T. Kanade and S. Baker, "Guide to the CMU Multi-PIE Database", *Carnegie Mellon University*, Technical report, 2007.

[18] D. G. Lowe, "Distinctive Image Features from Scale-invariant Key Points", *International Journal of Computer Vision*, 2004, vol. 60, pp. 91–110.

[19] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool, "A Comparison of Affine Region Detectors", *International Journal of Computer Vision*, 2005, vol. 65, pp. 43–72.

[20] R. Zass and A. Shashua, "Probabilistic Graph and Hypergraph Matching", *in IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[21] L. Torresani, V. Kolmogorov and C. Rother, "Feature Correspondence Via Graph Matching: Models and Global Optimization", *in European Conf. on Computer Vision*, 2008, pp. 596–609.

[22] C. K. Liu, A. Hertzmann and Z. Popovic, "Learning Physics-Based Motion Style with Nonlinear Inverse Optimization", *ACM Transactions on Graph*, 2005, vol. 24, pp. 1071–1081.

[23] E. Tola, V. Lepetit and P. Fua, "A Fast Local Descriptor for Dense Matching", *in Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[24] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors", *in Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2004, pp. 506–513.

[25] H. Bay, T. Tuytelaars and L. V. Gool, "SURF: Speeded Up Robust Features", *in European Conference on Computer Vision*, 2006, pp. 404–417.

[26] T. Cour, P. Srinivasan and J. Shi, "Balanced Graph Matching", *in Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada, 2006, pp. 313–320.

[27] F. De la Torre and M. H. Nguyen, "Parameterized Kernel Principal Component Analysis: Theory and Applications to Supervised and Unsupervised Image Alignment", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.