Facial feature localization using graph matching with higher order statistical shape priors and global optimization

Conference Paper · October 2010

DOI: 10.1109/BTAS.2010.5634502 · Source: IEEE Xplore

CITATIONS
14

READS
75

3 authors, including:

Shervin R. Arashloo
University of Surrey
27 PUBLICATIONS **381** CITATIONS

SEE PROFILE

William J Christmas
University of Surrey
74 PUBLICATIONS **1,225** CITATIONS

SEE PROFILE

Facial Feature Localization Using Graph Matching with Higher Order Statistical Shape Priors and Global Optimization

Shervin Rahimzadeh Arashloo*, Josef Kittler* and William J. Christmas*

***Abstract*— This paper presents a graphical model for deformable face matching and landmark localization under an unknown non-rigid warp. The proposed model learns and combines statistics of both appearance and shape variations of facial images (learnt purely from a set of frontal training images) in a complex objective function in an unsupervised manner. Local and global shape variations are included in the objective function as binary and higher order clique potentials. The proposed approach exploits the sparseness of facial features to reduce the complexity of inference over the probabilistic model. Besides presenting a method for face feature localization, the paper proposes a framework for incorporation of statistical shape priors as higher order cliques into MRFs.**

The problem of optimizing the objective function is performed using the dual decomposition approach in which the higher order subproblems based on point distribution models are formulated as instances of convex quadratic programs.

The evaluation of the approach for feature localization is performed both on the frontal and rotated images of the XM2VTS dataset images as well as images collected from Google. The method shows high robustness to partial occlusion, pose changes *etc.* The method is then applied as an initialization step for a more costly matching method and is shown to be instrumental in improving performance and reducing runtime.

I. INTRODUCTION

The idea of dividing an object into its constituting parts and modeling their spatial interaction is established as an effective approach for object modeling. In realistic situations such as in the presence of pose changes, background clutter, lighting and expression changes, partial occlusion, noise, resolution *etc.* which result in significant appearance changes between instances of the same object, methods based on this premise are commonly preferred [45] over holistic approaches [46] or part-based approaches which ignore configurational arrangements of object primitives [18]. The foregoing idea has been reflected in various applications, including image alignment for object recognition which can be considered as an integral part of all object recognition methods.

A popular framework for modeling conditional dependencies between object primitives is based on undirected graphical models, also known as Markov random fields. Very often, the dependencies between object primitives are limited to pairwise interactions to simplify the model and to reduce computational complexity. However, recent works have shown that priors of higher order can provide better

*The authors are with the Center for Vision, Speech and Signal Processing, University of Surrey, Guildford GU27XH, UK,{sr00048,j.kittler,w.christmas}surrey.ac.uk.

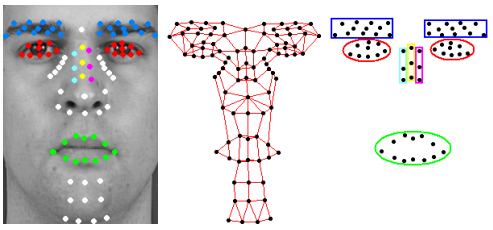


Fig. 1. From left to right: landmark points used for constructing the face graph superimposed on a sample face image, graph illustrating binary connectivities, higher order cliques used for different face components.

means to capture statistics of objects variations and hence result in superior performance in practice. Recently, the application of higher order priors has been the center of attention as recent advances in MRF optimization provide the necessary tools for incorporating such information into the hypothesized models. Although higher order priors have been known to be useful, their applicability in the context of MRFs has been limited due to the curse of dimensionality. Various approaches have recently addressed this issue and proposed different frameworks to incorporate a global prior into the model using various global optimization algorithms [49], [38], [27], [39].

In this work we present a graphical model using higher order shape priors for deformable face matching and feature localization. The proposed method learns the appearance and structure of faces in a probabilistic framework using a set of frontal training images in an unsupervised fashion and reduces the computational cost of MRF optimization by exploiting sparseness of facial features. Another novelty of the approach in the context of MRFs, is the incorporation of models of shape variation of the different components of face (*e.g.* eyes), based on point distribution models (PDM), as higher order clique potentials and formulate them as convex quadratic quadratic programming instances for which a variety of different approaches exist [7]. This is significantly important since it extends the application of higher order priors considered in specific forms [27], [38] to a more common and widely applied statistical model (PDM) for shape representation. While the application has been directed towards faces, the framework is more general and provides a principled way for incorporating such higher order statistical shape priors into graphical models. In the context of face recognition, the proposed methodology can be used in a variety of applications such as alignment of faces for geometric preprocessing, or initialization of other approaches like 3D morphable model and dense image matching methods

978-1-4244-7580-3/10/$26.00 ©2010 Crown

which are either computationally intractable or prohibitively expensive without proper initialization.

The paper is organized as follows: In section II we briefly review the literature on methods for face alignment. Section III explains the structure of the proposed model. In section IV the energy function combining texture and shape information is formulated. Section V discusses the approach we take for minimizing the energy. The experimental evaluation of the method on images collected from Google and on the rotation shots of XM2VTS database [34] is presented in section VI. In section VII conclusions are drawn.

II. Related Work

Object matching and alignment has been studied extensively *e.g.* in [11], [41], [26]. and face matching and alignment is no exception. One popular category is based on the statistical models built from a set of representative samples [19], [33], [36]. Two primary examples of such methods are active appearance models and their extensions [15], [26], [6] and 3D morphable model [12]. The matching is usually expensive and often needs a good initialization. Another drawback of 2D statistical models based on AAMs is the lack of generalization capability in extreme poses, in spite of the view-based versions [16] and the need for manual annotation of the training set. Other work in [14] employs a group-wise objective function to estimate non-rigid deformation. Other examples of the works based on shape constraints are [48], [40]. The work in [30] employs a component based discriminative approach using probabilistic shape constraints for face alignment. A similar approach to the current work is the elastic bunch graph matching (EBGM) approach [50]. Our approach differs from EBGM in various respects like node attributes, the geometric relations in the graphical structure and the optimization approach employed for inference. Also, the proposed method is completely unsupervised compared to EBGM for which manual intervention is needed for a couple of images. A graph-based method is also presented in [25].

Another interesting group of methods [22], [29], [44] employ a cost function defined as sum of entropies and a sequential method to find the transformation parameters. The work in [17] is an extension of previous approaches which uses a sum of squared error function for performing the alignment under the Lucas-Kanade's algorithm [32].

The work in [21] uses an elastic matching approach albeit discarding the relations between face image patches. Other works in [5] and [24] learn the changes in appearance of different patches of the faces using a training data.

III. Graph structure

Structural methods are based on the definition of a morphological model which is then combined with image measurements for object matching. In a sparse representation, the object is modeled using a number of landmark points. Then, given a statistical model, one learns from the training set independent and covariant probability distributions of the object appearance variations. These densities then enable one to describe the information contained in a new image based on the information observed in the training set. In practice, the training set is constructed by either manually labeling the landmarks for each instance of the object, or by inferring the landmark points by registering a labeled object with a number of selected landmark points to a set of non-labeled objects. In order to annotate our training set we used the method in [4] to register frontal images of 200 clients in the XM2VTS [34] dataset. The method in [4] is essentially a deformable image matching method which can establish pixel-wise correspondences between a pair of images. Annotation of the training set is hence performed by registering a labeled frontal face to every image in the database. In total we obtain a set of 1600 annotated images (8 images per each client) to train our model. It is worth reiterating that the proposed approach is completely *unsupervised* without any need for manual annotation of training images.

A. Selection of landmark points

Different approaches for representing and modeling faces use slightly different landmark points but the common characteristic is that the feature points are located around facial components *i.e.* eyes, eyebrows, nose, mouth *etc.* as in the active appearance models [15]. We discard the points on the contour of the face from our model since these regions lack distinctive features for matching. Assuming that important features of the face lie around the edges of facial components, after aligning a set of training images and averaging them, we chose a set of 92 landmark points based on the magnitudes of the edge map in the average face image as follows: 9 points for each eye, 12 points for each eyebrow, 12 points for the mouth, 10 points around the chin and finally 28 points for nose and surrounding regions. The set of adopted landmark points, superimposed on a sample face image, is illustrated in Fig.1.

B. Edges

Inclusion of an edge (connecting at most two nodes as opposed to hyper-edge) between any two nodes of the graph is executed manually based on the Euclidean distance. The aim is to ensure that a path exists between every two nodes of each component of face without the need for traversing from the nodes which do not belong to the same component. A graph illustrating the binary relations is depicted in Fig. 1 (middle).

C. Hyperedges

As noted earlier, although limiting the cardinality of cliques produces computational efficiency it may compromise the quality of the match. Point distribution models are useful for modeling shape variations. One can model the co-dependence of positions of the nodes jointly as in non-MRF based approaches [13]. In the MRF framework, this can be achieved by incorporating a hyper-edge containing all nodes of the graph, the potential of which is determined by the degree of deviation from the mean shape in the shape space

constructed using PCA. However, we do not used such a clique and make use of component-wise hyper-edges in our model for the following reasons:

- Variations in the shape of different components of face are almost independent of each other. In other words, representing deformations of all different parts of faces jointly using a unimodal distribution (*e.g.* gaussian) is neither realistic nor sufficient.
- From an optimization point of view, marginalizing over a very large clique including all nodes is computationally inefficient.

We also include a set of higher order priors in our model to constrain certain nodes on the nose to lie on straight lines.

IV. Energy Function

Matching the model to an image involves maximizing the *a posteriori probability* of observing the model given the image. In the MRF context, the *a posterior probability* is defined in terms of a Gibbs distribution and one usually minimizes the $-log$ of a posteriori probability of the model called the energy. Our energy function comprises different terms, representing different aspects of shape and texture variations. The energy function in the proposed model has the following form:

$$E(X;\theta) = \sum_{v \in V} \theta_v(x_v) + \sum_{c \in C} \theta_c(\mathbf{x_c}) \quad (1)$$

V corresponds to the set of nodes on the graph and C to the set of cliques including more than one node. θ_v and θ_c stand for unary and clique potentials of the graph. The cliques used in the current work have three different natures and cardinalities:

$$C = \{BI, L, PDM\} \quad (2)$$

BI represents binary relations, *L* represents third-order relations and *PDM* stands for higher than third-order cliques, discussed in the following.

A. Unary potentials

Different features for texture description exist in the literature such as SIFT [31], shape contexts [8], Gabor features [23], geometric blur [10]. Geometric blur features are known to be affine-invariant [10] and are used for shape matching and recognition [9]. For measuring texture similarities we use geometric blur features extracted from positive and negative edge channels in horizontal and vertical directions in an area of 35 pixels radius around the feature points. The circular area is sampled in 7 different radii and 18 different orientations.

Learning the main modes of texture variation

After extracting the geometric blur features for the 92 control points from our training set, we learn 92 probability density functions of texture variations. In the next step in order to remove redundancy and correlation effects and capturing main modes of texture variation we apply PCA to the set of features at each node separately. Then the similarity between each candidate point in the test image and each control point in our model is measured in the PCA feature space.

B. Pairwise potentials

The shape of an object can be modeled locally using binary relations. Using our training set, for each pair of nodes connected by an edge in the model graph, we estimate the mean and covariance of a 2D gaussian distribution for the relative positions of the nodes after aligning the training set using an affine transformation. We define the pairwise potentials of a pair of nodes in the graph in terms of the Mahalanobis distance from the mean configuration of the corresponding nodes:

$$\theta^{BI}_{(s,t)\in C}(x_s, x_t) = (d_{s,t}(x_s, x_t) - m_{s,t})^T C^{-1}_{s,t} (d_{s,t}(x_s, x_t) - m_{s,t}) \quad (3)$$

where $d_{s,t}(x_s, x_t)$ denotes the Euclidean distance between nodes s and t being assigned labels x_s and x_t. $m_{s,t}$ and $C^{-1}_{s,t}$ stand for the mean difference vector of coordinates and the inverse covariance matrix of their deviations obtained from the training set.

C. Higher-order potentials

In order to model the variations of face shape more accurately, we make use of point distribution models and include them in the face model as higher order cliques. These cliques capture higher order statistical variations of shapes of different components of face. In our model, we consider one such clique for each eye, one per each eyebrow, three third-order cliques on the nose and one for the mouth, Fig. 1. In the following we first describe the clique potentials based on point distribution models and then the third-order potentials used for the nose.

1) point distribution models: The shape of each facial component can be represented in a covariance space using a point distribution model. In order to construct a point distribution model for each component, we first align the training images using an affine transformation. The positions of all nodes contained in the clique under an affine transformation are used to estimate a mean shape for the corresponding component. Then in order to capture the main modes of shape variation, we apply PCA to the normalized positions of the nodes of the corresponding component in the training set. Then a configurational arrangement of a set of points in a facial component can be represented as:

$$A(\mathbf{x}_c) = \psi + \Phi \mathbf{w}(\mathbf{x_c}) \quad (4)$$

where $\mathbf{x}_c$ and ψ correspond to the configuration of nodes of the clique in test image and their mean shape coordinates, respectively. Φ is the matrix of M principal eigenvectors ($M < 2\times$cardinality of the clique) of the covariance matrix of the vectors of coordinates and $\mathbf{w}(\mathbf{x_c})$ is the vector of weights. A is a transformation, mapping the configuration (labels of nodes in the MRF) of nodes included in the clique into the corresponding spatial coordinates in the image frame under

an affine transformation. The clique potential is then defined as:

$$\theta_c^{PDM}(\mathbf{x}_c) = \mathbf{w}(\mathbf{x_c})^T \mathbf{w}(\mathbf{x_c}) \tag{5}$$

In practice one needs to minimize this potential in order to make the shape of the component under consideration as close as possible to those in the training set [51].

2) Linearity-based priors: The clique potentials of the third order on the nose are defined differently from other higher order potentials. For these cliques, since we have selected them in such a way that all the nodes in a clique lie on a straight line, we impose a prior representing an error function measuring the vertical offsets of the nodes from a line obtained by least square fitting. The above linearity assumption remains almost true even under pose and expression changes, since, compared to other facial components the nose is less deformable. In order to impose such priors we use regression to minimize the vertical offsets of points from the best fitted line. In this case the quality of the fit is given by:

$$\theta_c^L(\mathbf{x}_c) = \frac{{SS_{ij}}^2}{SS_{ii}SS_{jj}} \tag{6}$$

where the quantities SS_{ij}, SS_{ii} and SS_{jj} are given by:

$$SS_{ii} = \sum_{n=1}^{N} (i_n - \bar{i})^2 \tag{7}$$

$$SS_{jj} = \sum_{n=1}^{N} (j_n - \bar{j})^2 \tag{8}$$

$$SS_{ij} = \sum_{n=1}^{N} (i_n - \bar{i})(j_n - \bar{j}) \tag{9}$$

$N = 3$ and i and j denote the horizontal and vertical axes and i_n and j_n stand for the horizontal and vertical coordinates of the n^{th} point. $\bar{i}$ and $\bar{j}$ stand for the average of the coordinates of points in two directions.

V. Minimizing the energy using dual-decomposition approach

The idea of dual-decomposition method [28] is to decompose the original problem (the so-called master problem) into several easier MRF subproblems (the so-called slave MRFs) on each of which exact inference is tractable and then extracting a solution for the master by cleverly combining the solutions on the slaves which can be done based on an iterative projected subgradient approach. In other words, master acts as a coordinator which iteratively updates the costs of different configurations of each node in each subproblem separately so that the slaves agree on a common configuration at the end of the process.

The work in [27] extended the above framework from binary case to minimize functions of arbitrary arities. The generic optimizer for the higher-order MRFs proposed in [27] decomposes the master problem into several slaves in such a way that a separate sub-problem exists for each higher order clique.

In the decomposition approach, one requires $\hat{\theta} = \{\theta^\omega | \omega \in I\}$ to be a ρ-*reparameterization* of the original parameter vector θ [47] *i.e.*:

$$\sum_{\omega \in I} \rho_\omega \theta^\omega = \theta \tag{10}$$

Then for each subproblem a lower bound $\Phi_\omega(\theta^\omega)$ is defined which satisfies:

$$\Phi_\omega(\theta^\omega) \leq \min_x E(x; \theta^\omega) \tag{11}$$

It can be readily observed that the function

$$\Phi(\theta) = \sum_{\omega \in I} \rho_\omega \Phi_\omega(\theta^\omega) \tag{12}$$

is a lower bound of the original function in Eq. 1, *i.e.*

$$\Phi(\theta) \leq E(X^*; \theta) \tag{13}$$

where X^* is the optimal solution of Eq. 1. Following the same framework, we decompose the original energy in such a way that a separate subproblem exists for each higher order clique of facial components including only the nodes of the corresponding component. Binary relations are decomposed into two edge-disjoint spanning trees. In the following we describe how to solve each of these subproblems.

A. Higher-order subproblems

As noted earlier we associate one subproblem to each of our higher-order cliques. Two different kinds of higher order subproblems we use are solved as follows.

1) higher-order subproblems based on PDM: In an inference task, one needs to optimize a slave problem having as prior the distance defined in Eq. 5, hence the energy to minimize for each such clique is of the form:

$$\begin{aligned} E_c(\mathbf{x_c}) &= \theta_c^{PDM}(\mathbf{x}_c) + \sum_{v \in c} \theta_v(x_v) \\ &= \mathbf{w}(\mathbf{x}_c)^T \mathbf{w}(\mathbf{x}_c) + \sum_{v \in c} \theta_v(x_v) \end{aligned} \tag{14}$$

where we have included $\mathbf{x}_c$ to emphasize dependence of shape coefficients ($\mathbf{w}$) on it. Considering a clique of n nodes, $c = \{s_1, ..., s_n\}$, the problem of finding the optimum of the function in 14 can be defined as

$$\min_\mu \{\theta_c^{PDM} \mu(s_1; j_1) ... \mu(s_n; j_n) + \sum_{s;j} \theta_{s;j} \mu(s; j)\} \tag{15}$$

s.t.

$$\sum_j \mu(s; j) = 1$$

$$\mu(s; j) \in [0, 1].$$

In the following we show that the problem is in fact a *convex* quadratic programming problem. This then allows us to use convex optimization methods. Let Y denote an assignment, that is a vector of binary values $\{0, 1\}$ of dimension nL (n being the cardinality of the clique and L the number of admissible states for each node) which is obtained by concatenating n discrete potential functions (μ) of dimensionality L, each of which has all its components

zero except one component of value 1 indicating the state a node in the clique has been assigned. The problem of inferring a set of variables with a gaussian prior can be written in matrix from as [42]:

$$f(Y) = \frac{1}{2} Y^T H Y + BY \tag{16}$$

s.t.

$$AY = I_n$$

I_n is a vector of ones of dimension n. Under an affine transformation, matrix H and vector B are defined as:

$$\begin{aligned} H &= 2S^T M^T \Phi \Phi^T M S \\ B &= 2S^T M^T \Phi \Phi^T (t - \psi) + \theta_c^{PDM} \end{aligned} \tag{17}$$

where S is a matrix of dimensionality $2n \times nL$ mapping an assignment (Y) to its corresponding coordinates in the 2D image plane:

$$S = \begin{bmatrix} x_{1,1} & \dots & x_{1,L} & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ y_{1,1} & \dots & y_{1,L} & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & x_{2,1} & \dots & x_{2,L} & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & y_{2,1} & \dots & y_{2,L} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & x_{n,1} & \dots & x_{n,L} \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & y_{n,1} & \dots & y_{n,L} \end{bmatrix}$$

$x_{n,l}$ and $y_{n,l}$ denote horizontal and vertical coordinates of l^{th} candidate match for n^{th} node in the clique.

$M^{2n \times 2n}$ and $t^{2n \times 1}$ are the matrix (rotation and scale) and vector (translation) defining an affine transformation.

$$M = \begin{bmatrix} a & b & 0 & 0 & 0 & \dots & 0 & 0 \\ -b & a & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & a & b & 0 & \dots & 0 & 0 \\ 0 & 0 & -b & a & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \dots & a & b \\ 0 & 0 & 0 & 0 & 0 & \dots & -b & a \end{bmatrix} \tag{19}$$

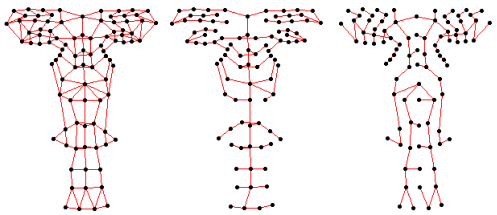
Assuming sf and α to be the scale factor and rotation angle respectively, $a = sf \times \cos\alpha$ and $b = sf \times \sin\alpha$

$$t = [t_x \ t_y \ t_x \ t_y \ \dots \ t_x \ t_y]^T \tag{20}$$

where t_x and t_y denote translations in two directions. It can be readily observed that matrix H in our problem is of the form DD^T hence positive semi-definite. Also, the points which satisfy the constraint form a convex set (any linear constraint defines a convex set). As a result, the quadratic program in Eq. 16 is a convex program. Quadratic programming is a well studied problem in nonlinear optimization field and many algorithms exist for optimization of such problems. In order to minimize the above function we use a method inspired by the work in [37] and use the following iterative algorithm in the primal space:

Consider node s and suppose that values $\mu(t;.)$ are fixed for all other nodes $t \neq s$, the optimal parameter $\mu(s;.)$ for node s is then given by:

$$\mu(s_1;.) = \arg\min_{\mu(s_1;.)} \{\theta_c^{PDM} \mu(s_1;j_1) \dots \mu(s_n;j_n) + \sum_j \theta_{s_1;j} \mu(s_1;j)\}$$



(a) Original graph:G (b) T^1 (c) T^2

Fig. 2. Decomposition of the original loopy graph into two edge-disjoint spanning trees: $edgeset_G = edgeset_{T^1} \cup edgeset_{T^2}$

subject to $\sum_j \mu(s;j) = 1$.

This can be obtained by taking:

$$j^*(s) = \arg\min_j \{\theta_c^{PDM} \mu(s_2;j_2) \dots \mu(s_n;j_n) + \theta_{s;j}\} \tag{21}$$

and then setting $\mu(s;j) = [j^* = j]$ where $[.]$ is one if its argument is true and zero otherwise. The method above is iterated until none of the node changes its label.

(18)The method is essentially a higher-order variant of the ICM (iterated conditional modes) approach. Theoretically the ICM algorithm terminates in a local minimum of the function. However because the objective function here is convex, a local minimum is the global minimum of the objective function. Similar observations have been made in other works [37], [20]. The drawback of such method is that the speed of convergence is dependent on the initialization conditions. In this work, we initialize the method using the configuration having minimum Euclidean distance to the mean shape which results in faster convergence.

2) Higher-order subproblems imposing linearity constraint: Solving these subproblems involves optimizing function of the form:

$$E_c(\mathbf{x_c}) = \theta_c^L(\mathbf{x}_c) + \sum_{v \in c} \theta_v(x_v) \tag{22}$$

Since the cardinalities of these cliques are not very large (L^3 states), we do an exhaustive search to find the optimum of these functions.

B. Binary subproblems

In order to solve these subproblems we decompose the graph containing at most pairwise potentials into two spanning trees (Fig. 2). Exact inference on these trees is performed using max-product algorithm. The goal is to compute single-node and joint pairwise min-marginals of the energy ($E_T(X')$) associated with the tree distribution:

$$\begin{aligned} \nu_s(x_s) &= \min_{\{X' | x'_s = x_s\}} E_T(X') \\ \nu_{st}(x_s, x_t) &= min_{\{X' | (x'_s, x'_t) = (x_s, x_t)\}} E_T(X') \end{aligned} \tag{23}$$

the computation of which is facilitated by performing only local computations in tree structured distributions and message passing between adjacent nodes [47].

C. Remarks

1) Interest points: In order to match the model to an image we first sample the image coarsely. We first select 1000 feature points based on edge magnitudes. This is then followed by a regular sampling of the whole image in a coarser scale (*e.g.* 1 sample in every block of size 10×10) if no sample already exist in the block under consideration. The texture similarity of each node in our model is then compared to the samples and the most similar 50 samples are considered as admissible states for that node.

2) Visibility assumption: In our experiments we have assumed that the feature points we are looking for are visible in the image and the model is forced to find a corresponding point for each node. Nevertheless, it is possible to include an occlusion label into the model along with a homogeneity constraint on the assignment of such label (as in [45]), if desirable.

3) Uniqueness constraint: Another point to address is the uniqueness constraint which basically means that two nodes of the model cannot be matched to the same position. In order to impose such a constraint into the model, one option is to use a *linear assignment* subproblem. By solving this problem (*e.g.* by using the Hungarian algorithm [2]) uniqueness constraint can be imposed on the linear subproblem and as a result in the optimal MAP solution.

4) Modeling rigid motion: Planar transformations are not the best choice for modeling the rigid motion of faces but they are simple and computationally efficient. An alternative can be to estimate local transformations for different parts of the face. However, we have not pursued this because in that case the estimated transformations would be more prone to errors before convergence and might cause the model to deviate from the true solution. In practice, we estimate an affine transformation for the whole model using Levenberg-Marquardt method along with RANSAC and refine the transformation as optimization proceeds. This transformation can then be used to update the binary relations according to the estimated transformation and make the model even more robust to global geometric transformations.

VI. Experimental evaluation

In this section we provide some experimental results first for matching the model to the different face images and illustrate the results. Next, we use the method in a verification scenario on the challenging rotation shots of the XM2VTS dataset [34].

A. Images taken from XM2VTS dataset

1) Frontal images: We first evaluate the performance of the method on frontal images of the XM2VTS dataset. Some examples are shown in Fig. 3. As it is evident from the results, the method can detect landmarks very accurately.

2) Partial occlusion due to beard and glasses: Next, we test the method against partial occlusion due to glasses and beard. As can be seen in Fig. 3, the method is quite robust to such occlusions and handles them very well even though some features are completely occluded due to beard.

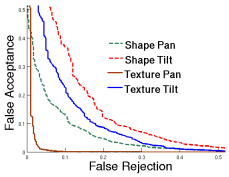


Fig. 5. ROC curves on the rotation shots of the Xm2VTS corpus.

TABLE I
Comparison of the current work to some other approaches

Method	This work	3D pose correction [43]	image matching [3]
EER	6.45	7.12	16.9

3) Pose variation: We also match the model to the rotation shots of the XM2VTS corpus. Some results are illustrated in Fig. 3. As can be seen from the examples, the model generalizes very good to non-frontal poses and performs reasonably well in the presence of severe pose changes.

B. Google image dataset

Next, we use the images collected from Google taken in real world conditions [21] for a cross database validation. Some examples are illustrated in Fig. 4. In the figure we have also included the results obtained by the CMU's face alignment system [19] for comparison. We have taken the results of this approach provided in [1] and test our model on the same set of images. It can be observed that the method generalizes very well across different databases and can handle various appearance changes, such as pose differences, outperforming the CMU's approach which fails in cases where in-depth rotation is present in the image. This is achieved in spite of the fact that we have trained our model using only frontal images. We were not able to compare our results with the method in [21] because we could not access the source code.

C. Face verification on the rotation shots of XM2VTS dataset

In the XM2VTS data set, the evaluation protocol is based on 295 subjects consisting of 200 clients, 25 evaluation imposters and 70 test imposters. Two error measures defined for a verification system are false acceptance and false rejection given below:

$$FA = EI/I * 100\%, \qquad FR = EC/C * 100\% \qquad (24)$$

where I is the number of imposter claims, EI the number of imposter acceptances, C the number of client claims and EC the number of client rejections. The performance of a verification system is often stated in *Equal Error Rate* (EER) in which the FA and FR are equal on the test set. In order to find dense correspondences between gallery and test images one may take different approaches such as using a regularized thin plate spline or a dense matching algorithm initialized by the landmarks located by the proposed model.

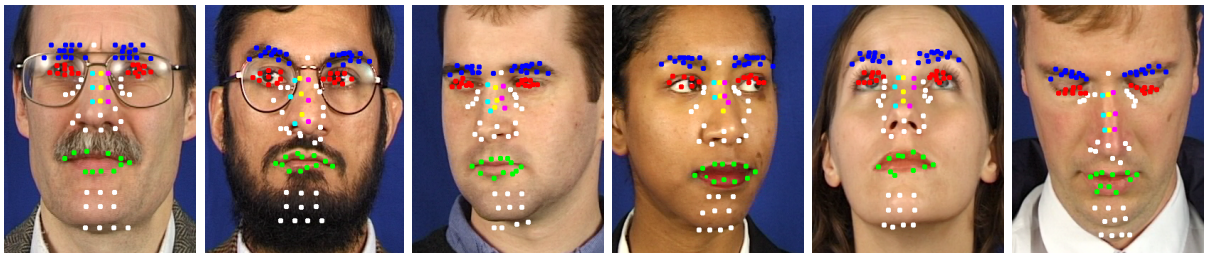
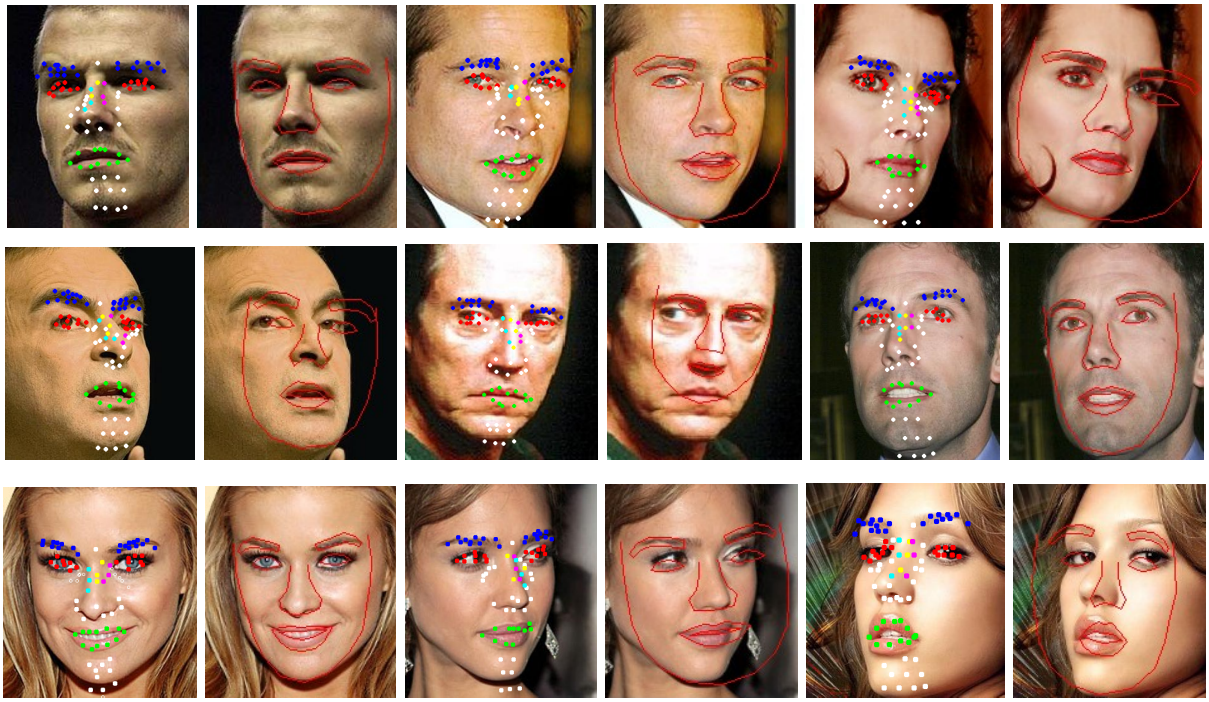


Fig. 3. Facial feature localization on the XM2VTS images partially occluded due to facial hair and glasses and rotated in-depth.



(a) Graph matching (b) CMU's method (c) Graph matching (d) CMU's method (e) Graph matching (f) CMU's method

Fig. 4. Results on the images collected from Google compared to CMU's method[19].

We take the second approach and employ the method in [4] initialized by the landmark points detected to establish pixel-wise correspondences between gallery and test images. Using the sparse matching as an initialization step, the runtime of the method in [4] reduces by a factor of 3. Histograms of uniform LBP patterns in a circular neighborhood [35] are used as texture features and compared the histograms using the χ^2 distance. For shape comparison, the deformation of a pair of faces is measured as proposed in [3]. The ROC curves using shape and texture information on the pan and tilt poses are illustrated in Fig. 5. The performance of the method is also compared to two other approaches in Table I. The one in [43] is based on a 3D pose normalization and the other in [3] is based on a similar matching method. One observes that the EER achieved in this work is lower than that of 3D pose normalization in [43]. The matching method in [3] is similar and hence the proposed sparse model initialization not only reduced the runtime but also the quality of the established match, resulting in superior performance.

VII. Conclusions and Future Works

The paper presented an MRF model for deformable face matching. The approach uses statistical models of texture and shape variations in building the model. The higher order statistical shape priors included in the graph, are shown to be instances of convex quadratic programs and solved by an iterative primal algorithm. The evaluation of the approach for feature localization is performed both on the XM2VTS dataset images and images collected from Google. The method shows high robustness to partial occlusion, pose changes *etc*. The method, used as an initialization step for a more costly dense matching approach, and is instrumental in improving performance and reducing runtime.

VIII. Acknowledgments

This work was supported by the project Mobile Biometry (MOBIO, www.mobioproject.org) grant IST-214324.

References

[1] http://www.vision.ee.ethz.ch/ zhuji/facealign/.

[2] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, February 1993.

[3] Shervin Rahimzadeh Arashloo and Josef Kittler. Pose-invariant face matching using mrf energy minimization framework. In *EMMCVPR*, pages 56–69, 2009.

[4] S.R. Arashloo and J.V. Kittler. Hierarchical image matching for pose-invariant face recognition. In *BMVC09*, 2009.

[5] A.B. Ashraf, S. Lucey, and Tsuhan Chen. Learning patch correspondences for improved viewpoint invariant face recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, 23-28 2008.

[6] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(10):1380 –1384, oct. 2004.

[7] Mokhtar S. Bazaraa, Hanif D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory And Algorithms*. Wiley-Interscience, May 2006.

[8] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, April 2002.

[9] Alexander C. Berg, Tamara L. Berg, and Jitendra Malik. Shape matching and object recognition using low distortion correspondence. In *In CVPR*, pages 26–33, 2005.

[10] Alexander C. Berg and Jitendra Malik. Geometric blur for template matching. In *CVPR (1)*, pages 607–614, 2001.

[11] A. Besbes, N. Komodakis, G. Langs, and N. Paragios. Shape priors and discrete mrfs for knowledge-based segmentation. pages 1295–1302, 2009.

[12] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, September 2003.

[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*, 61(1):38–59, January 1995.

[14] T. F. Cootes, C. J. Twining, V. Petrovic', R. Schestowitz, and C. J. Taylor. Groupwise construction of appearance models using piece-wise affine deformations. In *in Proceedings of 16 th British Machine Vision Conference*, pages 879–888, 2005.

[15] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *PAMI*, 23(6):681–685, Jun 2001.

[16] T.F. Cootes, K. Walker, and C.J. Taylor. View-based active appearance models. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 227 –232, 2000.

[17] M. Cox, S. Sridharan, S. Lucey, and J.F. Cohn. Least squares congealing for unsupervised alignment of images. pages 1–8, 2008.

[18] Chris Dance, Jutta Willamowski, Lixin Fan, Cedric Bray, and Gabriela Csurka. Visual categorization with bags of keypoints. In *ECCV*, 2004.

[19] Leon Gu and Takeo Kanade. A generative shape regularization model for robust face alignment. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 413–426, Berlin, Heidelberg, 2008. Springer-Verlag.

[20] Ken-Chung Ho. Iterated conditional modes for inverse dithering. *Signal Process.*, 90(2):611–625, 2010.

[21] Gang Hua and Amir Akbarzadeh. A robust elastic and partial matching metric for face recognition. In *International Conference on Computer Vision (ICCV)*, 2009.

[22] G.B. Huang, V. Jain, and E. Learned Miller. Unsupervised joint alignment of complex images. pages 1–8, 2007.

[23] Jarmo Ilonen, Joni-Kristian Kamarainen, Pekka Paalanen, Miroslav Hamouz, Josef Kittler, and Heikki Kälviäinen. Image feature localization by multiple hypothesis testing of gabor features. *IEEE Transactions on Image Processing*, 17(3):311–325, 2008.

[24] Takeo Kanade and Akihiko Yamada. Multi-subregion based probabilistic approach toward pose-invariant face recognition. In *In IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 954–959, 2003.

[25] D.R. Kisku, A. Rattani, M. Tistarelli, and P. Gupta. Graph application on face for personal authentication and recognition. In *Control, Automation, Robotics and Vision, 2008. ICARCV 2008. 10th International Conference on*, pages 1150 –1155, 17-20 2008.

[26] I. Kokkinos and A. Yuille. Unsupervised learning of object deformation models. In *ICCV 2007*, pages 1 –8, 14-21 2007.

[27] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order mrfs. pages 2985–2992, 2009.

[28] Nikos Komodakis, Nikos Paragios, and Georgios Tziritas. Mrf optimization via dual decomposition: Message-passing revisited. In *In ICCV*, 2007.

[29] E.G. Learned Miller. Data driven image models through continuous joint alignment. 28(2):236–250, February 2006.

[30] Lin Liang, Rong Xiao, Fang Wen, and Jian Sun. Face alignment via component-based discriminative search. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 72–85, Berlin, Heidelberg, 2008. Springer-Verlag.

[31] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[32] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI81*, pages 674–679, 1981.

[33] Iain Matthews and Simon Baker. Active appearance models revisited. *Int. J. Comput. Vision*, 60(2):135–164, 2004.

[34] K. Messer, J. Matas, J. Kittler, and K. Jonsson. Xm2vtsdb: The extended m2vts database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[35] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987, Jul 2002.

[36] Julien Pilet, Vincent Lepetit, and Pascal Fua. Fast non-rigid surface detection, registration and realistic augmentation. *Int. J. Comput. Vision*, 76(2):109–122, 2008.

[37] Pradeep Ravikumar and John Lafferty. Quadratic programming relaxations for metric labeling and markov random field map estimation. In *ICML*, pages 737–744. ACM Press, 2006.

[38] C. Rother, P. Kohli, Wei Feng, and Jiaya Jia. Minimizing sparse higher order energy functions of discrete variables. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1382–1389, 2009.

[39] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *CVPR 2006*, volume 1, pages 993 – 1000, 17-22 2006.

[40] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Face alignment through subspace constrained mean-shifts. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1034 –1041, sept. 2009.

[41] Thomas Schoenemann and Daniel Cremers. Globally optimal image segmentation with an elastic shape prior. In *ICCV*, pages 1–6, 2007.

[42] T.P. Speed and H.T. Kiiveri. Gaussian markov distributions over finite graphs. *The Annals of Statistics*, 14(138-150), 1986.

[43] JR Tena, RS Smith, M Hamouz, J Kittler, A Hilton, and J Illingworth. 2d face pose normalisation using a 3d morphable model. In *Proceedings of the International Conference on Video and Signal Based Surveillance*, pages 1–6, September 2007.

[44] Yan Tong, Xiaoming Liu, F.W. Wheeler, and P. Tu. Automatic facial landmark labeling with minimal supervision. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2097 –2104, 20-25 2009.

[45] Lorenzo Torresani, Vladimir Kolmogorov, and Carsten Rother. Feature correspondence via graph matching: Models and global optimization. In *ECCV '08*, pages 596–609. Springer-Verlag, 2008.

[46] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[47] M.J. Wainwright, T.S. Jaakkola, and A.S. Willsky. Map estimation via agreement on trees: message-passing and linear programming. *Information Theory, IEEE Transactions on*, 51(11):3697 – 3717, nov. 2005.

[48] Yang Wang, S. Lucey, and J.F. Cohn. Enforcing convexity for improved alignment with constrained local models. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1 –8, 23-28 2008.

[49] T. Werner. High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimisation (map-mrf). In *CVPR 2008*, pages 1–8, June 2008.

[50] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):775 –779, jul 1997.

[51] Zhong Xue, Stan Z. Li, and Eam Khwang Teoh. Ai-eigensnake: an affine-invariant deformable contour model for object matching. *Image Vision Comput.*, 20(2):77–84, 2002.