Viewpoint Invariant 3D Landmark Model Inference from Monocular 2D Images Using Higher-Order Priors

Chaohui Wang[1,2], Yun Zeng[3], Loic Simon[1], Ioannis Kakadiaris[4], Dimitris Samaras[3], Nikos Paragios[1,2]

[1]Center for Visual Computing, Ecole Centrale Paris, Châtenay-Malabry, France
[2]Equipe GALEN, INRIA Saclay - Ile-de-France, Orsay, France
[3]Department of Computer Science, Stony Brook University, NY, USA
[4]Computational Biomedicine Lab, University of Houston, TX, USA

Abstract

In this paper, we propose a novel one-shot optimization approach to simultaneously determine both the optimal 3D landmark model and the corresponding 2D projections without explicit estimation of the camera viewpoint, which is also able to deal with misdetections as well as partial occlusions. To this end, a 3D shape manifold is built upon fourth-order interactions of landmarks from a training set where pose-invariant statistics are obtained in this space. The 3D-2D consistency is also encoded in such high-order interactions, which eliminate the necessity of viewpoint estimation. Furthermore, the modeling of visibility improves further the performance of the method by handling missing correspondences and occlusions. The inference is addressed through a MAP formulation which is naturally transformed into a higher-order MRF optimization problem and is solved using a dual-decomposition-based method. Promising results on standard face benchmarks demonstrate the potential of our approach.

1. Introduction

3D model inference from 2D images is one of the most challenging problems in computer vision. This is due to the fact that both camera estimation and 3D model optimization have to be addressed within a single framework. In the most general case, the camera parameters are unknown, the 3D model itself usually inherits high complexity (high degrees of freedom even for non-articulated objects), while at the same time image features can be ambiguous, occluded and noisy. There are numerous applications involving the above scenario, such as traffic monitoring with 3D model based tracking [18], hand tracking [9], facial analysis [4] and medical imaging [17]. Such an inference process usually involves three steps: the first aims to determine a compact representation of the 3D model, the second to associate such a representation with the 2D image observation, and the last to recover the optimal parameters of the model.

Modeling variations of the 3D model requires a statistical parametric representation of the object of interest. Such representations in most cases are pose-variant, *i.e.*, all training examples are registered to a same referential frame where statistics are then built from the training data. One can cite active shape [8] and active appearance [7] models (ASMs and AAMs), which offer a good compromise between computational complexity and model expressiveness potential. Other representations adopt more complex statistical models that can vary from Mixtures of Gaussians (MoG) to non-parametric density functions [3]. In such a context, one has to address the curse of dimensionality (the dimensionality of the manifold to be learned versus the number of training samples).

Once the representation has been built, the next steps consist of defining an image likelihood and combining it with a 3D model prior towards optimal estimation of the 3D model. Since the image likelihood is related to both the 3D model configuration and the camera parameters, the model estimation is often achieved through an alternating search or EM-style approach [14]. Given an initial 3D-2D correspondence map, the camera parameters are first estimated and then used to define the fitting error between the model and the image. This error is to be optimized by gradient-driven methods and iterative search processes so as to estimate both the correspondences and the optimal model configuration. Despite the promising performance of such a scheme, the fact that explicit estimation of the camera viewpoint parameters is required in the process is a major drawback, since coordinate-descent approaches are prone to be trapped in local minima and provide no guarantee on the optimality of the estimations.

Graphical models have become a dominant approach in computer vision and have been employed to address a num-

ber of vision problems (*e.g.*, [5, 19, 20]), which is mainly due to their strength in terms of the quality of the optimum. The use of higher-order models has raised more and more attention (*e.g.*, [15, 12, 22]), along with the development of efficient optimization methods. Higher-order interactions can naturally introduce invariance to certain class of transformations like translation/rotation/scale, while at the same time they can deal with vision tasks of important complexity. The objective of our paper is to take benefit of their strength and propose a unified formulation to estimate 3D models from 2D images without alternating search.

The main contribution of this paper is a probabilistic inference approach that does not require explicit viewpoint estimation, while being able to jointly optimize the pose parameters and the corresponding landmarks selection as well as explicitly handling missing correspondences and occlusions via a visibility modeling. To this end, we formulate the problem as a maximum a posteriori (MAP) estimation task which involves 3D pose parameters, associated 2D correspondences and visibility states. We derive a posterior probability as the product of an image likelihood, a visibility prior, a 3D geometric prior and a projection consistency prior constraining the 2D and 3D configurations. In order to circumvent the need of viewpoint estimation, we adopt a high-order decomposition of the 3D model that enables to determine the projection error between a given 3D configuration and the corresponding 2D landmark positions in a distributed manner. Furthermore, an explicit visibility modeling is also introduced to cope with misdetections and outliers. The MAP inference is then naturally transformed into a higher-order MRF optimization problem and all the latent variables are inferred using a one-shot optimization over a factor graph [3] through dual-decomposition [2, 16, 20, 22]. The proposed formulation has been validated in the context of 3D facial pose estimation from 2D images. Promising results on standard face benchmarks demonstrate the potential of our method.

The remainder of this paper is organized as follows. In Sec. 2, we present the probabilistic formulation for the joint estimation of the 3D pose, its visibility states and the 2D correspondences. The individual likelihoods with respect to geometry, 3D-to-2D consistency, visibility, and image support are presented in Sec. 3 while the corresponding higher-order graphical model is discussed in Sec. 4. Experimental results compose Sec. 5, while discussion and future work conclude the paper in Sec. 6.

2. Probabilistic 3D-2D Inference Framework

We consider a point-distribution shape model composed of a set $\mathcal{V}$ of landmarks located on the surface of the 3D object of interest. Let latent variable $X_i = (X_i^{(3)}, X_i^{(2)})$ denote the 3D and 2D positions of a landmark i ($i \in \mathcal{V}$). More specifically, $X_i^{(3)}$ and $X_i^{(2)}$, 3-dimensional and 2-dimensional vectors respectively, denote the 3D position of landmark i in the model space and the 2D position in the observed image $\mathbf{I}$. Each variable X_i takes a value x_i from its possible configuration set $\mathcal{X}_i = \mathcal{X}_i^{(3)} \times \mathcal{X}_i^{(2)}$, where $\mathcal{X}_i^{(3)}$ and $\mathcal{X}_i^{(2)}$ denote the 3D and 2D position candidate sets, respectively. Due to the fact that landmarks may be invisible, we also introduce a visibility variable O_i for landmark i [19]. $O_i = 1$ when the landmark is visible in the 2D image space, and $O_i = 0$ otherwise.

Given the observed image $\mathbf{I}$, the estimation of the 3D-2D positions and the visibility of the landmarks is formulated as a maximization of the posterior probability of $(\mathbf{X}, \mathbf{O}) = ((X_i)_{i\in\mathcal{V}}, (O_i)_{i\in\mathcal{V}})$ over their domains $\mathcal{X} = \prod_{i\in\mathcal{V}} \mathcal{X}_i$ and $\mathcal{O} = \{0, 1\}^{|\mathcal{V}|}$:

$$(\mathbf{x}, \mathbf{o})^{\text{opt}} = \underset{(\mathbf{x},\mathbf{o})\in\mathcal{X}\times\mathcal{O}}{\arg\max}\ p(\mathbf{x}, \mathbf{o}|\mathbf{I}) \tag{1}$$

The posterior probability $p(\mathbf{x}, \mathbf{o}|\mathbf{I})$ is:

$$\begin{aligned} p(\mathbf{x}, \mathbf{o}|\mathbf{I}) &= p(\mathbf{x}, \mathbf{o}, \mathbf{I})/p(\mathbf{I}) \\ &\propto p(\mathbf{x}, \mathbf{o}, \mathbf{I}) \\ &= p(\mathbf{I}|\mathbf{x}, \mathbf{o}) \cdot p(\mathbf{x}, \mathbf{o}) \\ &= p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{o}) \cdot p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o}) \cdot p(\mathbf{o}|\mathbf{x}^{(3)}) \cdot p(\mathbf{x}^{(3)}) \\ &= \underbrace{p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o})}_{\text{Image Likelihood}} \cdot \underbrace{p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o})}_{\text{Projection Prior}} \cdot \underbrace{p(\mathbf{o})}_{\text{Visibility Prior}} \cdot \underbrace{p(\mathbf{x}^{(3)})}_{\text{3D Model Prior}} \end{aligned} \tag{2}$$

where $p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o})$ encodes the likelihood of the observed image given the 2D position configurations $\mathbf{x}^{(2)}$ and the visibility states $\mathbf{o}$ of the landmarks, $p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o})$ encodes the projection prior from the 3D configuration $\mathbf{x}^{(3)}$ to the 2D configuration of the landmarks, $p(\mathbf{o})$ denotes the visibility prior on the landmarks, and $p(\mathbf{x}^{(3)})$ denotes the prior on the 3D configurations of the landmarks.

Note that this probabilistic formulation can be directly applied to the estimation of 3D (or 2D) configuration of the landmarks given 2D (or 3D) configuration, simply by instantiating the variables whose configurations are known.

3. Definitions of the Probability Terms

In this section, we define all the probability terms which are involved in the posterior probability $p(\mathbf{x}, \mathbf{o}|\mathbf{I})$ (Eq. 2).

3.1. Image Likelihood

The image likelihood $p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o})$ measures the occurrence probability of the observed image $\mathbf{I}$, given the 2D position configurations $\mathbf{x}^{(2)}$ and the visibility states $\mathbf{o}$ of the landmarks. If we assume, without loss of generality, that the landmarks are independent in terms of appearance, then we can define $p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o})$ as follows:

$$p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o}) \propto \prod_{i\in\mathcal{V}} p(\mathbf{I}|x_i^{(2)}, o_i) \tag{3}$$

Regarding $p(\mathbf{I}|x_i^{(2)}, o_i)$, there are two cases:

1. When $O_i = 1$, the landmark's position is informative and $p(\mathbf{I}|x_i^{(2)}, o_i)$ denotes the likelihood of the observed image given that landmark i is located at position $x_i^{(2)}$, which can be defined using the output of a classifier such as Randomized Forest [6].

2. When $O_i = 0$, the landmark's position is not informative and $p(\mathbf{I}|x_i^{(2)}, o_i)$ denotes a uniform distribution, thus we assume that $p(\mathbf{I}|x_i^{(2)}, o_i) = \hat{p}$ (constant).

3.2. Projection Prior

The projection prior $p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o})$ measures the occurrence possibility of the 2D positions $\mathbf{x}^{(2)}$ of the landmarks when the 3D positions $\mathbf{x}^{(3)}$ and the visibility states $\mathbf{o}$ are given, which is modeled using Gibbs distribution:

$$p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o}) \propto \exp\{-\frac{f(\mathbf{x}, \mathbf{o})}{T}\} \tag{4}$$

where T is temperature, and the energy function $f(\mathbf{x}, \mathbf{o})$ encodes inconsistency between the 3D and 2D configurations of the landmarks taking the visibility states into account (the smaller $f(\mathbf{x}, \mathbf{o})$ is, the better is the correspondence between $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(2)}$).

Without loss of generality, we use the weak-perspective camera configuration [1] to model the projection from 3D points to 2D points[1]. Let us first consider a triplet $\mathbf{t} \in \mathcal{T} = \{\mathbf{t}|\mathbf{t} \subseteq \mathcal{V} \text{ and } |\mathbf{t}| = 3\}$ of landmarks that are all visible. Their 3D-2D positions $\mathbf{x_t}$ determine at most two projection mappings $\mathbf{P}_{\mathbf{x_t}}^{(s)}$ ($s \in \{1, 2\}$) [1, 11] corresponding to two reflective symmetric camera configurations. Then for any additional visible point i, we can measure the error $e_{\mathbf{x_t}}(x_i)$ between its 2D position $x_i^{(2)}$ and the value obtained by projecting its 3D position $x_i^{(3)}$, *i.e.*:

$$e_{\mathbf{x_t}}(x_i) = \min_{s \in \{1,2\}} \left\| \mathbf{P}_{\mathbf{x_t}}^{(s)}(x_i^{(3)}) - x_i^{(2)} \right\| \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean norm, and between the two feasible projections we consider the most prominent one with respect to the considered 2D configuration [1]. On the contrary, if one or more of these four points are invisible, we set a constant energy $\hat{E}$ as the projection error $e_{\mathbf{x_t}}(x_i)$, which can be understood as an upper bound of the average projection error which is allowed between four points. Therefore, we define the error function $e_{\mathbf{x_t},\mathbf{o_t}}(x_i, o_i)$ by taking the visibility states into account as:

$$e_{\mathbf{x_t},\mathbf{o_t}}(x_i, o_i) = w_{\mathbf{t}} \cdot \begin{cases} e_{\mathbf{x_t}}(x_i) & \text{if } o_j = 1, \forall j \in \mathbf{t} \cup \{i\} \\ \hat{E} & \text{otherwise} \end{cases} \tag{6}$$

[1]In the proposed framework, the weak-perspective camera model can be easily replaced by other camera models such as the perspective model.

where $w_{\mathbf{t}}$ is a confidence weight for the error measure obtained under the mapping determined by the positions of the points in clique $\mathbf{t}$, which will be presented later in this section. And then, the 3D-2D consistency between a quadruplet $\mathbf{c}$ of landmarks consists of the sum of the errors which are determined by taking all possible combinations of triplets within the quadruplet and evaluating the projection error on the remaining point:

$$e(\mathbf{x_c}, \mathbf{o_c}) = \sum_{\mathbf{t} \subset \mathbf{c}} e_{\mathbf{x_t},\mathbf{o_t}}(x_{\mathbf{c}\setminus\mathbf{t}}, o_{\mathbf{c}\setminus\mathbf{t}}) \tag{7}$$

Finally, we define the energy function $f(\mathbf{x}, \mathbf{o})$ as the sum of $e(\mathbf{x_c}, \mathbf{o_c})$ over all the quadruplet, *i.e.*:

$$f(\mathbf{x}, \mathbf{o}) = \sum_{\mathbf{c} \in \mathcal{C}} e(\mathbf{x_c}, \mathbf{o_c}) \tag{8}$$

where $\mathcal{C} = \{\mathbf{c}|\mathbf{c} \subseteq \mathcal{V} \text{ and } |\mathbf{c}| = 4\}$ denotes the set of all quadruplets.

Last, we should note that we can further combine other cues in this projection prior, such as regional texture similarity.

Robust Confidence Weight

Since the projection matrix estimation is unstable when considering triplets of 3D points that are nearly collinear [1], we introduce a confidence weight $w_{\mathbf{t}}$ to modulate the error contribution of each triplet of points. For a triangle $\Delta_{\mathbf{x_t}^{(3)}}$ consisting of a triplet $\mathbf{t}$ of points with 3D positions $\mathbf{x_t}^{(3)}$, we define the non-collinear coefficient $\text{NC}(\mathbf{x_t}^{(3)})$ using the square root of its area $\text{Area}(\Delta_{\mathbf{x_t}^{(3)}})$ and its perimeter $\text{Perim}(\Delta_{\mathbf{x_t}^{(3)}})$ as follows:

$$\text{NC}(\mathbf{x_t}^{(3)}) = \frac{2 \times 3^{\frac{3}{4}} \times \text{Area}^{\frac{1}{2}}(\Delta_{\mathbf{x_t}^{(3)}})}{\text{Perim}(\Delta_{\mathbf{x_t}^{(3)}})} \tag{9}$$

We can observe that $\text{NC}(\mathbf{x_t}^{(3)}) = 1$ for an equilateral triangle and $\text{NC}(\mathbf{x_t}^{(3)}) = 0$ when the three points are collinear. Then we learn the confidence weight $w_{\mathbf{t}}$ by averaging the non-collinear coefficients for each triplet $\mathbf{t}$ over the training data:

$$w_{\mathbf{t}} = \frac{1}{M} \sum_{m=1}^{M} \text{NC}(\mathbf{x}_{\mathbf{t},m}^{(3)}) \tag{10}$$

where M denotes the number of training samples.

Specification of the Projection Error

Regarding the computation of $e_{\mathbf{x_t}}(x_i)$, we use the efficient method proposed in [1] to compute directly the projection of a 3D point under the projection determined by a triplet

of corresponding 3D-2D points without calculating the projection mapping. We refer readers to [1] for more details.

Collinear triplets of points lead to degenerate configurations from which we cannot obtain a solution for the projection mapping. In this case, the corresponding error term $e_{\mathbf{x_t}}(x_i)$ in Eq. 5 is not well-defined. To deal with this, we consider two different scenarios: (i) When we have a prior knowledge that the 3D positions of a triplet $\mathbf{t}$ of points have to be collinear, we simply ignore the corresponding error measure by defining $e_{\mathbf{x_t}}(x_i) = 0$ (this is consistent with the confidence weight defined in Eq. 10, *i.e.*, $w_{\mathbf{t}} = 0$ leads to zero contribution to $f(\mathbf{x})$); (ii) Otherwise, we define $e_{\mathbf{x_t}}(x_i) = +\infty$ if $\mathbf{x}_{\mathbf{t}}^{(3)}$ are collinear so that the final solution of $\mathbf{x}_{\mathbf{t}}^{(3)}$ cannot be exactly collinear. By doing so, the term $e_{\mathbf{x_t}}(x_i)$ is well-defined for all the cases. For the sake of clarity, hereafter, we assume that the definition of $e_{\mathbf{x_t}}(x_i)$ in Eq. 5 implicitly includes the definition in the degenerate case.

3.3. Visibility Prior

We introduce the visibility variable $\mathbf{O}$ to achieve a more precise modeling of the 3D-2D estimation, due to the fact that a landmark can be invisible. The notion of "invisibility" encodes occlusions and self-occlusions in the 3D space, as well as misdetection due to insufficient image support or classification failure.

The inference process is performed by considering, for each landmark i, a number of 2D positions which lead to the highest probabilities $p(\mathbf{I}|x_i^{(2)})$ towards composing the set of plausible solutions for $\mathcal{X}_i^{(2)}$, expecting that at least one candidate is (or close to) the true position. However, because of erroneous detection or occlusions, it is possible that all the candidates are far from the ground truth. In such a context, we define the notion of "visibility" as whether the true 2D correspondence of the landmark is captured by the candidate set. More specifically, $O_i = 1$ means that at least one candidate in $\mathcal{X}_i^{(2)}$ is close to the ground truth, and $O_i = 0$ stands for the opposite case.

The prior probability $p(\mathbf{o})$ is defined as follows:

$$p(\mathbf{o}) = \prod_{i \in \mathcal{V}} p(o_i) \tag{11}$$

where $p(o_i)$ denotes the prior probability of the visibility of each individual landmark i and is modeled as a Bernoulli distribution $\text{Bern}(o_i|\mu_i)$ with parameter $\mu_i = \Pr(O_i = 1)$. In practice, it is usually reasonable to assume the same parameter $\mu > 0.5$ for all the landmarks [20].

3.4. 3D Model Prior

The training data are used to learn a 3D shape model. No assumption on registration between surfaces is being made. However, we assume that correspondences have been determined for the landmarks among the samples of the training set. The key concern of shape modeling is how to capture the inherent variability of the class of objects from a reasonable small training set using a compact representation that can be easily adopted towards an efficient inference. We adopt the pose-invariant prior in [22] which is based on the relative lengths of a triplet of points, and extend it in a more general formulation where the cliques can be of any higher order. Such a prior does not require the estimation of the global pose in the training and testing stages and eliminates the bias caused by such estimations.

Let us consider a clique[2] $\mathbf{c}$ ($\mathbf{c} \subseteq \mathbf{V}$ and $|\mathbf{c}| \geq 3$) of landmarks, we enumerate all the pairs $\mathcal{P}_{\mathbf{c}} = \{(i,j)|i,j \in \mathbf{c} \text{ and } i < j\}$ of points. Let $d_{ij} = \left\| x_i^{(3)} - x_j^{(3)} \right\|$ denote the Euclidean distance between points i and j ($(i,j) \in \mathcal{P}_{\mathbf{c}}$). We obtain the relative distance $\hat{d}_{ij}$ by normalizing the distance d_{ij} over the sum of the distances between the pairs of points involved in clique $\mathbf{c}$, *i.e.*,

$$\hat{d}_{ij} = d_{ij} / \sum_{(i,j) \in \mathcal{P}_{\mathbf{c}}} d_{ij} \tag{12}$$

Since for clique $\mathbf{c}$, any relative distance $\hat{d}_{ij}$ is a linear combination of the others (*i.e.*, $\sum_{(i,j) \in \mathcal{P}_{\mathbf{c}}} \hat{d}_{ij} = 1$), we store all the relative distances, except one in a vector $\hat{\mathbf{d}}_{\mathbf{c}} = (\hat{d}_{ij})_{(i,j) \in \bar{\mathcal{P}}_{\mathbf{c}}}$, where $\bar{\mathcal{P}}_{\mathbf{c}}$ contains the pairs that are involved in the vector $\hat{\mathbf{d}}_{\mathbf{c}}$. Then, the statistics on $\hat{\mathbf{d}}_{\mathbf{c}}$ are learned from the training data. We can model its distribution $p_{\mathbf{c}}(\hat{\mathbf{d}}_{\mathbf{c}})$ using standard probabilistic models such as MoGs and Parzen-Windows. Finally, we define the prior probability of the 3D configuration as:

$$p(\mathbf{x}^{(3)}) \propto \prod_{\mathbf{c} \in \mathcal{C}} p_{\mathbf{c}}(\hat{\mathbf{d}}_{\mathbf{c}}(\mathbf{x}_{\mathbf{c}}^{(3)})) \tag{13}$$

where $\mathcal{C}$ denotes the set of all cliques, and $\hat{\mathbf{d}}_{\mathbf{c}}(\mathbf{x}_{\mathbf{c}}^3)$ denotes the mapping from the 3D position $\mathbf{x}_{\mathbf{c}}^3$ of the clique $\mathbf{c}$ to the relative distance vector $\hat{\mathbf{d}}_{\mathbf{c}}$.

4. Higher-order MRF Formulation

The data likelihood, the 3D-2D consistency, the visibility prior and the 3D shape model, presented in Sec. 3, can be naturally encoded within a higher-order MRF model where latent variables are to be inferred through an energy minimization. In this perspective, the negative logarithm of the posterior probability (Eq. 2) is factorized into the potentials of the MRF and constitutes the MRF energy.

To this end, we use a node to model a landmark i ($i \in \mathcal{V}$) with its latent 3D-2D position X_i and its visibility O_i.

[2] As presented in Sec. 4, we use 4-order cliques (quadruplets) in this work, *i.e.*, $|\mathbf{c}| = 4$. However, other higher-order cliques $\mathbf{c}$ ($|\mathbf{c}| \geq 3$) can also be used in this shape model.

Actually, we can use a single random variable[3] to encode X_i and O_i compactly by simply defining a special label "occ" within 2D position candidate set $\mathcal{X}_i^{(2)}$ such that:

$$x_i = \begin{cases} (x_i^{(3)}, x_i^{(2)}) & \text{if } O_i = 1 \\ (x_i^{(3)}, \text{occ}) & \text{if } O_i = 0 \end{cases} \tag{14}$$

This compact representation is valid because the 2D position $x_i^{(2)}$ is meaningless when the landmark i is occluded (*i.e.*, when $O_i = 0$, the image likelihood $p(\mathbf{I}|\mathbf{x}^{(2)}, \mathbf{o})$ and the projection prior $p(\mathbf{x}^{(2)}|\mathbf{x}^{(3)}, \mathbf{o})$ are constant with respect to $x_i^{(2)}$.).

In order to factorize the potential functions, we use a fourth-order clique to model a quadruplet $\mathbf{c}$ of landmarks. Due to the bijective mappings between nodes and landmarks and between fourth-order cliques and quadruplets, we reuse $\mathcal{V}$ and $\mathcal{C}$ to denote the node set and the clique set which determine the topology of the MRF. The 3D and 2D positions of the landmarks are estimated through the minimization of the MRF energy $E(\mathbf{x})$:

$$\mathbf{x}^{\text{opt}} = \arg\min_{\mathbf{x} \in \mathcal{X}} E(\mathbf{x}) \tag{15}$$

Here, the energy of the MRF is defined as the negative logarithm of the posterior probability in Eq. 2 (up to an additive constant) and can be factorized into the following form:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} U_i(x_i) + \sum_{\mathbf{c} \in \mathcal{C}} H_c(\mathbf{x_c}) \tag{16}$$

where $\mathbf{x_c}$ denotes the configuration $(x_i)_{i \in \mathbf{c}}$ of clique $\mathbf{c}$.

Singleton potential $U_i(x_i)$ $(i \in \mathcal{V})$ encodes the data likelihood (see section 3.1) and the visibility prior (see section 3.3). After taking the negative logarithm, we obtain its definition as follows:

$$U_i(x_i) = \begin{cases} -\log p(\mathbf{I}|x_i^{(2)}) & \text{if } x_i^{(2)} \neq \text{"occ"} \\ \lambda_1 & \text{if } x_i^{(2)} = \text{"occ"} \end{cases} \tag{17}$$

where λ_1 is a constant coefficient.

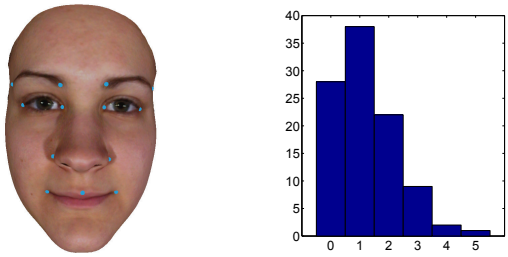
Higher-order clique potential $H_{\mathbf{c}}(\mathbf{x_c})$ $(\mathbf{c} \in \mathcal{C})$ is defined as follows:

$$H_{\mathbf{c}}(\mathbf{x_c}) = \lambda_2 \cdot H_{\mathbf{c}}^{(1)}(\mathbf{x_c}) + \lambda_3 \cdot H_{\mathbf{c}}^{(2)}(\mathbf{x_c}) \tag{18}$$

where $\lambda_2 > 0$ and $\lambda_3 > 0$ are two balancing constants, $H_{\mathbf{c}}^{(1)}(\mathbf{x_c})$ encodes the 3D statistic geometry constraints implied by the shape prior on the 3D configuration of the landmarks, and $H_{\mathbf{c}}^{(2)}(\mathbf{x_c})$ encodes the 3D-2D projection prior:

$$\begin{cases} H_{\mathbf{c}}^{(1)}(\mathbf{x_c}) = -\log p_{\mathbf{c}}(\hat{\mathbf{d}}_{\mathbf{c}}(\mathbf{x}_{\mathbf{c}}^{(3)})) \\ H_{\mathbf{c}}^{(2)}(\mathbf{x_c}) = e(\mathbf{x_c}, \mathbf{o_c}(\mathbf{x_c})) \end{cases} \tag{19}$$

[3]In order to reduce the number of symbols used, we reuse X_i to denote this new random variable. Accordingly, we reuse x_i, $\mathcal{X}_i$ and the other related notations.



(a) (b)

Figure 1. (a) The distribution of landmarks; (b) The histogram presenting the distribution of the number of missing 2D correspondences in the first experiment.

where $\mathbf{o_c}(\mathbf{x_c})$ denotes the binary visibility values that are recovered from $\mathbf{x_c}$ using Eq. 14, and the definitions of $e(\mathbf{x_c}, \mathbf{o_c})$ and $p_{\mathbf{c}}(\alpha(\mathbf{x}_{\mathbf{c}}^{(3)}))$ have been presented in Sec. 3.2 and 3.4, respectively.

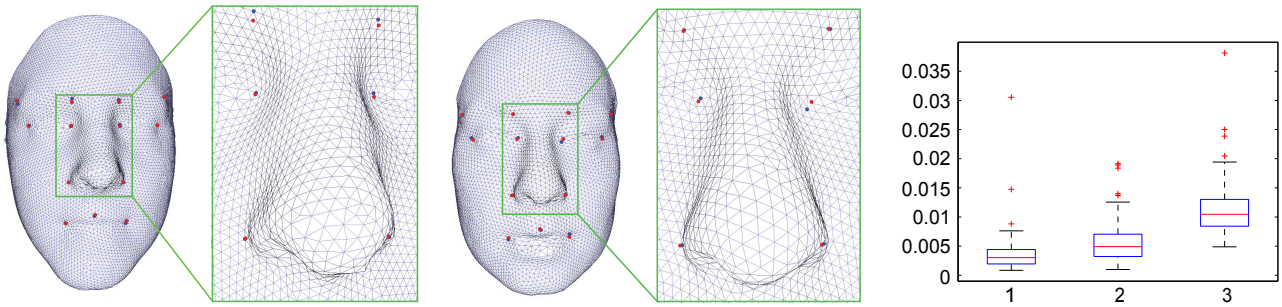
Dual-Decomposition MRF Inference: Regarding the inference of the proposed higher-order MRF, we adopt the dual-decomposition optimization framework [2, 16], which is considered to be the state-of-the-art towards MAP-MRF inference [16, 20] in particular when handling higher-order MRFs [15, 22]. Based on this framework, we decompose the original problem which is difficult to solve directly into a set of sub-problems which can be solved very efficiently. The solutions of the sub-problems are combined using projected subgradient method [16, 20] to achieve the solution of the original problem. Regarding the decomposition, like [22], we decompose the original graph into a set of factor trees which can be solved within polynomial time using max-product belief propagation algorithm [3].

5. Experimental Results

5.1. Experimental Settings

The performance of the proposed method was evaluated on the publicly-available facial expression datasets *BU-3DFE* [24] and *BU-4DFE* [23]. The former consists of 3D range data of 6 prototypical facial expressions of 100 different subjects (56 female and 44 male), and the latter is composed of 3D dynamic facial expressions of 101 different subjects (58 female and 43 male). The subjects included in both datasets are of various ethnic/racial origins.

The considered model consists of 13 landmarks (eyes, nose, mouth and eyebrows as shown in Fig. 1(a)). In the inference stage, its 3D initialization was done by randomly picking one training example. Regarding the 3D positions of the landmarks, the search was guided by a coarse-to-fine scheme and sparse sampling strategy in a similar way as [13]. Upon convergence of the algorithm, we performed *Procrustes Analysis* [10] to obtain the similarity transform



(a) (b) (c)

Figure 2. Results of the first experiment. (a) and (b): 3D model estimation results. In each sub-figure, 3D face mesh is provided for measuring visually the error between the resulting positions (in red) of landmarks and the ground truth (in blue). (c): Boxplots for the distributions of dissimilarity measures for qualitatively evaluating the 3D model estimation. c.1: Results obtained by the proposed method; c.2: Results obtained by the version without visibility modeling; c.3: Initialization of the model. On each box, the central mark in red is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

between the estimated 3D model and the ground truth, then transformed the estimated one into the referential frame of the ground truth. In terms of quantitative evaluation, a common goodness-of-fit criterion is the squared error standardized by the scale of the object. Thus, *Procrustes distance* [10] was used as the dissimilarity measure E^{d} to evaluate our method quantitatively, which can be computed as follows:

$$E^{\mathrm{d}} = \sum_{i\in\mathcal{V}} \left\| \dot{x}_i^{(3)} - \hat{x}_i^{(3)} \right\|^2 / \sum_{i\in\mathcal{V}} \left\| \hat{x}_i^{(3)} - \hat{\mathbf{C}}^{(3)} \right\|^2 \qquad (20)$$

where $\dot{x}_i^{(3)}$ and $\hat{x}_i^{(3)}$ denote the resulting and ground truth 3D positions of landmark i, respectively, $\hat{\mathbf{C}}^{(3)} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} \hat{x}_i^{(3)}$ is the center of the ground truth model. The smaller E^{d} is, the closer the resulting model is to the ground truth.

In all the experiments, the concept of leave-one-out cross-validation was adopted towards the evaluation of the method. In this context, we do the validation on a sample while using the remaining samples as training data, and such a validation is done for all the samples contained in a dataset using the same parameter settings. Regarding the 3D model prior (Eq. 13), we modeled the probability distribution $p_{\mathbf{c}}(\hat{\mathbf{d}}_{\mathbf{c}}(\mathbf{x}_{\mathbf{c}}^{(3)}))$ between a quadruplet $\mathbf{c}$ of points using a two-component Gaussian Mixture.

5.2. Qualitative Results and Quantitative Analysis

First, we considered 100 samples of the neutral expression from *BU-3DFE*, one from each subject. The 2D landmark correspondence space was associated with 5 labels, four corresponding to the 2D position candidates and the last to the occlusion label "occ". On top of the ground truth correspondence, noise was added to generate erroneous 2D candidates as well. Furthermore, for 10% of the landmarks (randomly sampled), the true correspondence was removed and replaced with a random position in the image plane, which produced between 0 and 5 missing 2D correspondences for each test (see Fig. 1(b)). Figs. 2(a) and (b) present 3D model esitmation results. Fig. 2(c).3 and Fig. 2(c).1 (*i.e.*, the boxes 3 and 1 in Fig. 2(c)) depict the statistics of the dissimilarity measure E^{d} (Eq. 20) for the initialization and the resulting 3D model obtained by the proposed method, respectively. The qualitative and quantitative evaluations demonstrate that our method leads to well-estimated 3D models even when correspondences are partially missing. Furthermore, in order to demonstrate the impact of the visibility modeling, we have also evaluated an alternative version (without visibility modeling) of the proposed method where the "occ" label was removed from the 2D candidate set of each node, and show the obtained statistics of E^{d} in Fig. 2(c).2. Based on the comparison of Fig. 2(c).1 and Fig. 2(c).2, we can conclude that the visibility modeling indeed leads to significantly better performance.

Second, we employed the facial feature point detector proposed in [21] to obtain the 2D position candidates for 101 samples of *BU-4DFE*, also one from each subject. Such a detector is based on Gabor features and boosting classifiers, and can well localize the considered landmarks from observed 2D images (Figs. 3(a)-(f)), though errors may still be present in some tests. We also performed a leave-one-out cross-validation in this experiment. Figs. 3(a')-(f') show six 3D model estimation results of different qualities and Fig. 3(g) presents the statistics of E^{d} for the proposed method and the version without visibility modeling. These results further demonstrate the potential of the proposed method to infer the 3D configuration of the model from 2D observed images with misdetections/occlusion handling.

Last but not least, we compared our method with an al-

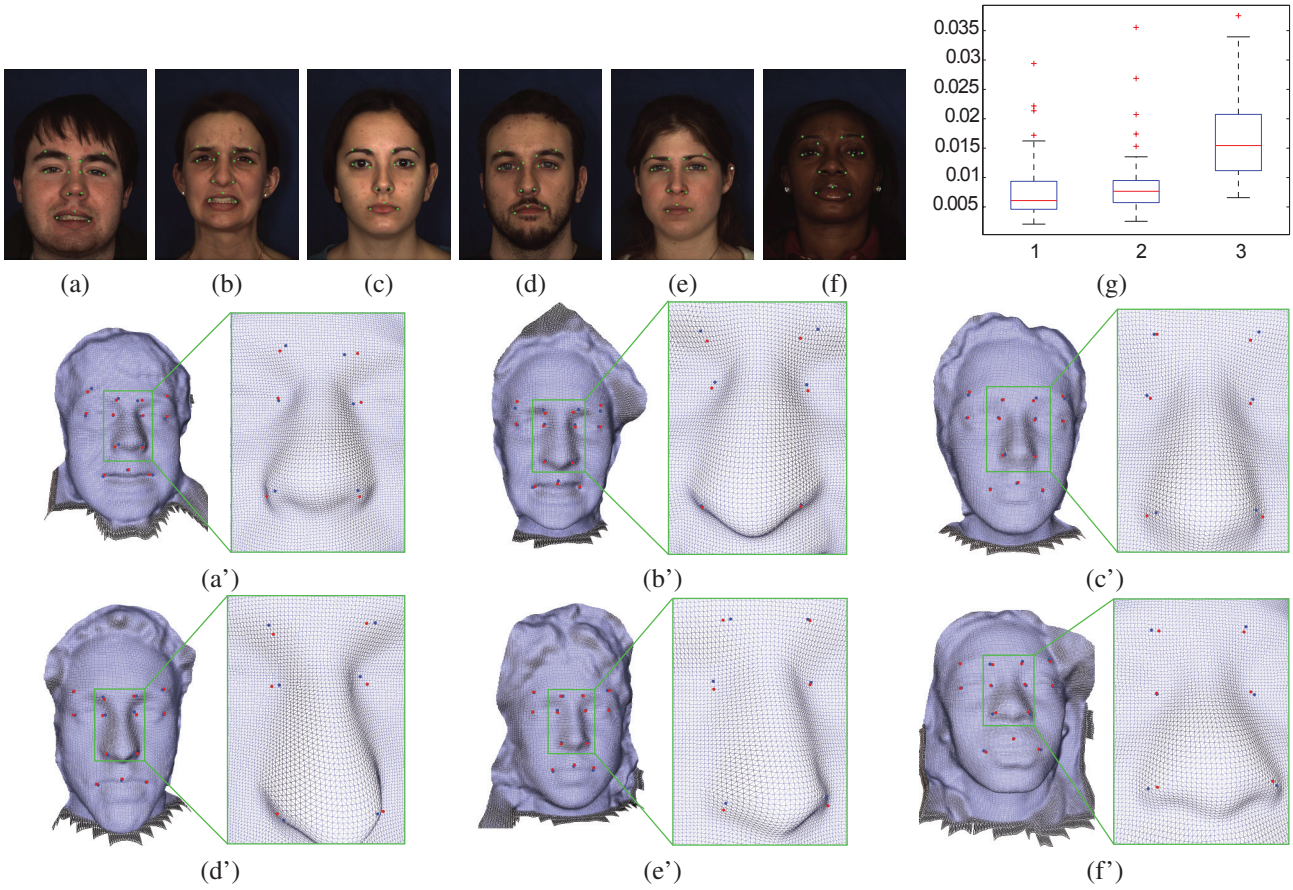


Figure 3. Results of the second experiment. (a)-(f): 2D landmark detection results [21]; (a')-(f'): The corresponding 3D model estimation results. (g): Boxplots for the distributions of dissimilarity measures for qualitatively evaluating the 3D model estimation. g.1: Results obtained by the proposed method; g.2: Results obtained by the version without visibility modeling; g.3: Initialization of the model.

ternative method (ASM+RANSAC) with a relaxed condition where we assumed that the ground truth 2D correspondences were known. For each test, we first learned an ASM [8] from the training data. Then, we used RANSAC [11] to estimate the camera projection function based on the initialization of the shape model and the given ground truth 2D correspondences. Once the projection function was estimated, we searched for the best shape configuration by minimizing the errors between the projections of the 3D points and their 2D correspondences. Furthermore, we evaluated both methods using two different initializations: besides the "random sample" initialization used throughout the experiments, we also tested the "mean-shape" initialization where we chose one example as the reference, registered all the other training examples to it and computed the mean shape as initialization. We performed leave-one-out cross-validation on all the 2500 samples of *BU-3DFE* dataset and the quantitative evaluation is shown in Fig. 4. Figs. 4.1 and 4.4 show that our method performed equally well with the two different initializations, which demonstrates robustness with respect to the choice of initialization. The evaluation of ASM+RANSAC is presented in Figs. 4.2 and 4.5. We observe from Fig. 4 that the dissimilarity measure of our method is approximately 3 to 5 times lower compared to ASM+RANSAC, which demonstrates that our method performs significantly better than ASM+RANSAC and is highly robust with respect to the initialization.

In conclusion, the results of all the experiments demonstrate that our method, despite the important variability of pose and facial geometry, has well estimated the 3D configuration of the model even with the existence of misdetections, and outperforms significantly the alternative methods.

6. Conclusion

In this paper, we have introduced a novel approach for 3D landmark model inference from a monocular 2D view that combines the estimation of the 3D pose, the visibility states and the 2D correspondences. The main innovations of the method are the absence of camera parameters estimation, the ability to model geometric consistency through local priors, the explicit modeling of visibility and

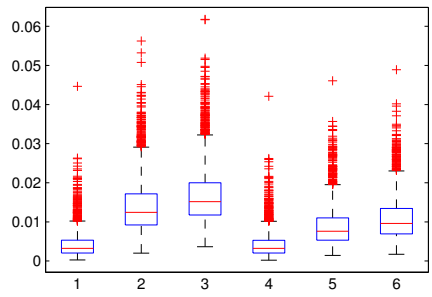


Figure 4. Comparison with ASM+RANSAC in terms of dissimilarity measures. 1. Our method with random-sample initialization; 2. ASM+RANSAC with random-sample initialization; 3. The random-sample initialization; 4. Our method with mean-shape initialization; 5. ASM+RANSAC with mean-shape initialization; 6. The mean-shape initialization.

the one-shot optimization to jointly infer all the variables. We have evaluated our method on standard facial datasets with promising results.

Future work concerns first the achievement of a better model decomposition towards recovering the smallest subset of higher-order interactions that can express the 3D geometric manifold, which could drastically decrease the computational complexity of the method. The use of more advanced parameterizations of the manifold which go beyond simple 3D landmark positions (*e.g.*, the entire surface through some kind of local interpolation) would open new application domains of our method like body pose estimation or medical image analysis where 2D partial acquisition of 3D objects is frequent. Last but not least, faster optimization algorithms of higher-order MRFs, including potential implementations of existing optimizers on GPUs, could be beneficial to our approach both in terms of the considered application as well as in terms of modularity with respect to other 3D pose estimation problems from 2D images.

Acknowledgments

This work was partially supported by the European Research Council Starting Grant DIOCLES (ERC-STG-259112) and the MEDICEN Competitive Cluster sterEOS+ grant.

References

[1] T. D. Alter. 3-D pose from 3 points using weak-perspective. *IEEE TPAMI*, 16(8):802–808, 1994.

[2] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, 1999.

[5] Y. Boykov and G. Funka-Lea. Graph cuts and efficient N-D image segmentation. *IJCV*, 70(2):109–131, 2006.

[6] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

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

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models - their training and application. *CVIU*, 61(1):38–59, 1995.

[9] M. de La Gorce, N. Paragios, and D. J. Fleet. Model-based hand tracking with texture, shading and self-occlusions. In *CVPR*, 2008.

[10] I. L. Dryden and K. V. Mardia. *Statistical shape analysis*. John Wiley & Sons Inc., 1998.

[11] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[12] B. Glocker, H. Heibel, N. Navab, P. Kohli, and C. Rother. TriangleFlow: Optical flow with triangulation-based higher-order likelihoods. In *ECCV*, 2010.

[13] B. Glocker, N. Komodakis, G. Tziritas, N. Navab, and N. Paragios. Dense image registration through MRFs and efficient linear programming. *Medical Image Analysis*, 12(6):731–741, 2008.

[14] L. Gu and T. Kanade. 3D alignment of face in a single image. In *CVPR*, 2006.

[15] N. Komodakis and N. Paragios. Beyond pairwise energies: Efficient optimization for higher-order MRFs. In *CVPR*, 2009.

[16] N. Komodakis, N. Paragios, and G. Tziritas. MRF optimization via dual decomposition: Message-passing revisited. In *ICCV*, 2007.

[17] R. Kurazume, K. Nakamura, T. Okada, Y. Sato, N. Sugano, T. Koyama, Y. Iwashita, and T. Hasegawa. 3D reconstruction of a femoral shape using a parametric model and two 2D fluoroscopic images. *CVIU*, 113(2):202–211, 2009.

[18] D. Roller, K. Daniilidis, and H.-H. Nagel. Model-based object tracking in monocular image sequences of road traffic scenes. *IJCV*, 10(3):257–281, 1993.

[19] E. B. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky. Distributed occlusion reasoning for tracking with nonparametric belief propagation. In *NIPS*, 2004.

[20] L. Torresani, V. Kolmogorov, and C. Rother. Feature correspondence via graph matching: Models and global optimization. In *ECCV*, 2008.

[21] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *IEEE SMC*, 2005.

[22] C. Wang, O. Teboul, F. Michel, S. Essafi, and N. Paragios. 3D knowledge-based segmentation using pose-invariant higher-order graphs. In *MICCAI*, 2010.

[23] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3D dynamic facial expression database. In *Int. Conf. on Automatic Face and Gesture Recognition*, 2008.

[24] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3D facial expression database for facial behavior research. In *Int. Conf. on Automatic Face and Gesture Recognition*, 2006.