**Face Alignment by Explicit Shape Regression**

Xudong Cao · Yichen Wei · Fang Wen · Jian Sun

**Abstract**

We present a very efﬁcient, highly accurate, “Explicit Shape Regression” approach for face alignment. Unlike previous regression-based approaches, we directly learn a vectorial regression function to infer the whole facial shape (a set of facial landmarks) from the image and explicitly minimize the alignment errors over the training data. The inherent shape constraint is naturally encoded into the regressor in a cascaded learning framework and applied from coarse to ﬁne during the test, without using a ﬁxed parametric shape model as in most previous methods. To make the regression more effective and efﬁcient, we design a two-level boosted regression, shape indexed features and a correlation-based feature selection method. This combination enables us to learn accurate models from large training data in a short time (20 min for 2,000 training images), and run regression extremely fast in test (15 ms for a 87 landmarks shape). Experiments on challenging data show that our approach signiﬁcantly outperforms the state-of-the-art in terms of both accuracy and efﬁciency.

**Keywords**

Face alignment · Shape indexed feature · Correlation based feature selection · Non-parametric shape constraint · Tow-level boosted regression

**1 Introduction**

Face alignment or locating semantic facial landmarks such as eyes, nose, mouth and chin, is essential for tasks like face recognition, face tracking, face animation and 3D face modeling. With the explosive increase in personal and web photos nowadays, a fully automatic, highly efﬁcient and robust face alignment method is in demand. Such requirements are still challenging for current approaches in unconstrained environments, due to large variations on facial appearance, illumination, and partial occlusions.

A face shape S = [x1 , y1 , ..., xNfp , yNfp ]Tconsists of Nfp facial landmarks. Given a face image, the goal of face alignment is to estimate a shape S that is as close as possible to the true shape S, i.e., minimizing

||S – S^||2 . (1)

The alignment error in Eq. (1) is usually used to guide the training and evaluate the performance. However, during testing, we cannot directly minimize it as S is unknown. According to how S is estimated, most alignment approaches can be classiﬁed into two categories: optimization-based and regression-based.

Optimization-based methods minimize another error function that is correlated to (1) instead. Such methods depend on the goodness of the error function and whether it can be optimized well. For example, the AAM approach (Matthews and Baker 2004; Sauer and Cootes 2011; Saragih and Goecke 2007; Cootes et al. 2001) reconstructs the entire face using an appearance model and estimates the shape by minimizing the texture residual. Because the learned appearance models have limited expressive power to capture complex and subtle face image variations in pose, expression, and illumination, it may not work well on unseen faces. It is alsowell known that AAM is sensitive to the initialization due to the gradient descent optimization.

Regression-based methods learn a regression function that directly maps image appearance to the target output. The complex variations are learnt from large training data and testing is usually efﬁcient. However, previous such methods (Cristinacce and Cootes 2007; Valstar et al. 2010; Dollar et al. 2010; Sauer and Cootes 2011; Saragih and Goecke 2007) have certain drawbacks in attaining the goal of minimizing Eq. (1). Approaches in (Dollar et al. 2010; Sauer and Cootes 2011; Saragih and Goecke 2007) rely on a parametric model (e.g., AAM) and minimize model parameter errors in the training. This is indirect and sub-optimal because smaller parameter errors are not necessarily equivalent to smaller alignment errors. Approaches in (Cristinacce and Cootes 2007; Valstar et al. 2010) learn regressors for individual land-marks, effectively using (1) as their loss functions. However, because only local image patches are used in training and appearance correlation between landmarks is not exploited, such learned regressors are usually weak and cannot handle large pose variation and partial occlusion.

We notice that the shape constraint is essential in all methods. Only a few salient landmarks (e.g., eye centers, mouth corners) can be reliably characterized by their image appearances. Many other non-salient landmarks (e.g., points along face contour) need help from the shape constraint—the correlation between landmarks. Most previous works use a para-metric shape model to enforce such a constraint, such as PCA model in AAM (Cootes et al. 2001; Matthews and Baker 2004) and ASM (Cootes et al. 1995; Cristinacce and Cootes 2007).

Despite of the success of parametric shape models, the model ﬂexibility (e.g., PCA dimension) is often heuristically determined. Furthermore, using a ﬁxed shape model in an iterative alignment process (as most methods do) may also be suboptimal. For example, in initial stages (the shape is far from the true target), it is favorable to use a restricted model for fast convergence and better regularization; in late stages (the shape has been roughly aligned), we may want to use a more ﬂexible shape model with more subtle variations for reﬁnement. To our knowledge, adapting such shape model ﬂexibility is rarely exploited in the literature.

In this paper, we present a novel regression-based approach without using any parametric shape models. The regressor is trained by explicitly minimizing the alignment error over training data in a holistic manner—all facial landmarks are regressed jointly. The general idea of regressing non-parametric shape has also been explored by Zhou and Comaniciu (2007).

Our regressor realizes the shape constraint in an non-parametric manner: the regressed shape is always a linear combination of all training shapes. Also, using features across the image for all landmarks is more discriminative than using only local patches for individual landmarks. These properties enable us to learn a ﬂexible model with strong expressive power from large training data. We call our approach “Explicit Shape Regression”.

Jointly regressing the entire shape is challenging in the presence of large image appearance variations. We design a boosted regressor to progressively infer the shape—the early regressors handle large shape variations and guarantee robustness, while the later regressors handle small shape variations and ensure accuracy. Thus, the shape constraint is adaptively enforced from coarse to ﬁne, in an automatic man-ner. This is illustrated in Fig. 1 and elaborated in Sect. 2.3.

Our explicit shape regression framework is inspired by the cascaded pose regression proposed by Dollar et al. (2010). In their work, a sequence of random fern regressors are learnt to predict the object pose parameters progressively. In each iteration, image features not only depend on the image content, but also depend on the predicted pose parameter from last iteration. Such pose-indexed features provide better geometric invariance and greatly enhance the regressor’s performance. In their experiment, this method has also been used to estimate face shape which is modeled by a simple parametric ellipse (Dollar et al. 2010).

Our method improves the cascaded pose regression framework in several important aspects and works better for face alignment problem. We adopt a non-parametric representation, directly estimate the facial landmarks by minimizing the alignment error instead of parameter error. Consequently, the underlying shape constraint is preserved automatically. To address the very challenging high-dimensional regression problem, we further propose several improvements: a two-level boosted regression, effective shape indexed features, a fast correlation-based feature selection method and sparse coding based model compression so that: (1) we can quickly learn accurate models from large training data (20 min on 2,000 training samples); (2) the resulting regressor is extremely efﬁcient in the test (15 ms for 87 facial landmarks); (3) the model size is reasonably small (a few megabytes) and applicable in many scenarios. We show superior results on several challenging datasets.